DeepGeoVision: Deep Learning Techniques for Various Remote Sensing Vision Tasks

Thesis submitted to the
Indian Institute of Technology Guwahati
for the award of the degree

of

Doctor of Philosophy

in

Computer Science and Engineering

Submitted by

Avinash Kumar Chouhan

Under the guidance of

Prof. Arijit Sur

Department of Computer Science and Engineering

Indian Institute of Technology Guwahati

December 6, 2024

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Prof. Arijit Sur
Department of Computer Science and Engineering
IIT Guwahati
Email : arijit@iitg.ac.in

Certificate

This is to certify that this thesis entitled, **"DeepGeoVision: Deep Learning Techniques for Various Remote Sensing Vision Tasks"**, being submitted by **Avinash Kumar Chouhan**, to the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, for partial fulfillment of the award of the degree of Doctor of Philosophy, is a bonafide work carried out by him under my supervision and guidance. The thesis, in my opinion, is worthy of consideration for award of the degree of Doctor of Philosophy in accordance with the regulation of the institute. To the best of my knowledge, it has not been submitted elsewhere for the award of the degree.

.....
Prof. Arijit Sur
Professor

Date: December 6, 2024
Place: Guwahati

Declaration

I certify that:

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor.
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor is not in a position to check for any possible instance of plagiarism within this submitted work.

Date: December 6, 2024
Place: Guwahati

(Avinash Kumar Chouhan)

Dedicated to

My Family

For their unconditional love, blessings, support, and constant inspiration

Acknowledgements

I am deeply grateful to Prof. Arijit Sur for the invaluable opportunity to work as a research scholar under his mentorship. His guidance and expertise have been instrumental in shaping my PhD journey. I also appreciate the freedom he granted me to explore and select my research topics. My sincere thanks go to the PhD admission interview committee of the CSE department for their confidence in my research potential, which has played a crucial role in my academic progress.

I am indebted to the esteemed seniors of the Multimedia Lab, particularly Dr. Brijesh and Dr. Satisha, for their invaluable guidance throughout my research journey and to lab mates Dr. Anirban, Mr. Alik, Mr. Sandeepan, Mr. Sonal, Mr. Akshay, Mr. Suklav, Mr. Utkarsh and Dr. Kanan. Special recognition is owed to lab senior Dr. Prasen for his unwavering support and encouragement during periods of adversity, especially amidst the challenges of paper rejections. A heartfelt acknowledgment goes to my dear colleagues at IIT Guwahati, Dr. Alakesh, Dr. Nilotpal, and Dr Maithelee, whose support and guidance have been instrumental in my academic endeavors.

I express my sincere appreciation to the earlier Director NESAC, Mr. PLN Raju, for entrusting me with the opportunity to pursue my PhD and to my superior, Dr. Dibyajyoti Chutia, for inspiring me to delve into the realm of Deep Learning for remote sensing applications. Additionally, I am very grateful to the current Director NESAC, Dr. Shiv Prasad Aggarwal, for fostering an environment conducive to AI-based research, which has fueled my enthusiasm for swift progress.

My deepest gratitude is reserved for my parents and brother, Mr. Basant Kumar Chouhan, Mrs. Shusheela Chouhan, and Mr. Abhishek Kumar Chouhan, whose unwavering support and inspiration have been the cornerstone of my journey. I am also thankful to my parents-in-law, Mr. Sadananda Kalita and Mrs. Binita Kalita, my brother-in-law, Mr. Nikunja Kalita, and my sister-in-law, Mrs. Deepika Kalita, for their encouragement and assistance throughout my PhD coursework.

Finally, I am immensely grateful to my wife, Dr. Mamita Kalita, whose unwavering belief in my potential and steadfast support have motivated me throughout

this journey.

Abstract

The development of intelligent automated vision applications is a need of the hour in the remote sensing domain due to the exponential growth in data generated by aerial and satellite sensors. This dissertation proposes deep learning models for four important Remote Sensing Vision Task (RSVT)s: Semantic Segmentation (SS), Change Detection (CD), Image Translation (IT), and Unsupervised Domain Adaptation (UDA) in four contributory chapters, respectively. The first contributory chapter shows that the problem of Semantic Segmentation (SS) can be modeled as an image-to-image mapping where pixel-level classification is required. Pixel level classification is challenging for the High Resolution (HR) aerial images due to the presence of the tiny objects in low-frequency, and more information details for such tiny objects are required for dense semantic labeling. Encoder-decoder based architectures are very common in the recent literature for aerial image SS. However, these architectures often suffer from information loss due to the up-and-down sample process. In the first contributory chapter of this dissertation, two deep models named Aggregated Context Network (ACNet) and HybridNet are proposed for effective SS. In ACNet, we proposed a variant of encoder-decoder-based architecture where residual attentive skip connections are incorporated to reduce information loss. We also added a multi-context block in each of the encoder units to capture the multi-scale and multi-context features and use dense connections for effective feature extraction. HybdridNet extends a HR network with dense connection integration to preserve the original resolution and better parameter sharing. We also incorporate a lightweight self-attention module for positional attention, which results in better segmentation maps. Additionally, it uses a generalized hough transform-based deep voting module for pixel dependencies extraction.

Change Detection (CD) in satellite images is an important research area as it has a wide range of applications in natural resource monitoring, geo-hazard detections, urban planning, etc. Identifying physical changes on the ground and avoiding spurious changes due to other reasons like co-registration issues, changes in illumination conditions, sun angle, and presence of cloud and fog is a challenging task. Moreover, CD improves with error-free image reconstruction as an auxiliary task, which needs joint feature representation by a feature extractor. In the second contributory chapter of this dissertation, we proposed two deep models (Difference image Reconstruction enhanced Multiresolution Network (DRMNet), Triad Multitask Learning for change detection (TMLNet)) to resolve issues for the recent existing CD methods. In DRMNet, we proposed a multitask learning based CD model where two parallel pipeline architectures predict change map and image difference. The proposed model takes two images and their difference as input and provides them to a Backbone Network (BN). The output of the BN is fed into the proposed Multi-Scale Attention Module

(MSAM) for the effective identification of changes in multi-temporal and very HR aerial images. In another parallel path, the output of the BN is down-sampled and passed to the proposed Deconvolution with a Sub-pixel Convolution Module (DSCM) to generate image difference. Two loss functions are utilized in two parallel paths to train the overall model in an end-to-end supervised setting. In another model for CD (TMLNet), we proposed a novel triad (which is a combination of input images and its difference) learning-based multiresolution architecture TMLNet for effective CD. We also proposed a multi-context local self-attention module to efficiently calculate long-range pixel relations with multiple contexts. An enhanced backbone module with top-down connections and multi-scale channel and spatial attention is utilized to generate change maps. It provides less noisy features extracted through the backbone. Laplacian pyramid loss is used to preserve small details in feature reconstruction.

In the third contributory chapter, we consider Image Translation (IT) as a vision application that converts input to output, which may be of another domain or resolution. Two models named Encoder Decoder based Conditional GAN (EDCGAN) and Multi Context Dense Network (MCDNet) are proposed for IT, one for Synthetic Aperture Radar (SAR) to optical translation and the other for multi-frame super-resolution. HR optical images are heavily utilized in various remote sensing applications. The optical images cannot reflect the actual ground information in cloudy conditions. SAR images are used to solve this for their ability to see through clouds. However, SAR images are usually available with coarser resolutions. So, there is a need to produce an optical image from a SAR image to overcome bad weather and poor resolution in a single go. In this chapter, we have proposed a SAR to RGB IT method called EDCGAN, which is an encoder-decoder-based conditional GAN that uses multi-scale attentive discrimination to get accurate SAR to RGBIT. In addition, we have used residual connections and spatial & channel-wise attention for better feature representation. In the second work of this chapter, we consider Multi-frame Super Resolution (SR), which utilizes multiple low-resolution images to generate a single high-resolution image. Multi-frame SR methods face difficulty in handling spatial and temporal dependencies of pixels. In this work, a model named MCDNet is proposed to handle spatial and temporal pixel dependencies using multiple approaches of global average pooling, multiple size kernels, and self-attention.

In the fourth and final contributory chapter, we observed that Unsupervised Domain Adaptation (UDA) is an important task that transfers learned knowledge from the source domain to the unseen target domain. Domain shift is the major challenge faced by UDA methods, which is caused by differences in appearance, distribution, decision boundary, sensor platforms, capturing conditions, etc. The goal of UDA methods is to minimize the existing shift between domains. Domain invariant feature representation is essential for effective UDA. Mix representations combine inputs of both domains, which are help-

ful for the generation of domain-consistent features using self-training. Domain alignment at multiple levels is required for effective adaptive segmentation between domains. Considering the above-mentioned issues, this chapter proposes two domain adaptation models named Masked Domain Adversarial Adaptation Network (MDAANet) and Reconstruction Assisted Domain Adaptation (RADA) for domain adaptive segmentation and nighttime tracking. MDAANet contains a masked domain dual adaptation approach, joint adversarial alignment, consistency enforcement, and feature dissimilarity-based alignment for effective UDA. It allows the adaptation in input, feature, and output levels, producing better UDA for SS. It increased the inter-class dissimilarities and attained input-output level combined adaptation. In the second work, we consider the problem of nighttime Object Tracking (OT). We found that despite the impressive progress made by numerous state-of-the-art trackers on large-scale datasets, visual OT at nighttime remains challenging because of low light (brightness) conditions, lack of contrast, very low variability among feature distributions, etc. In addition, the lack of paired (labeled) data for nighttime tracking makes it infeasible for supervised learning-based modeling. As a countermeasure, unsupervised domain adaptation based tracking can resolve this issue. In this work, we have proposed RADA with adversarial learning for nighttime OT for nighttime tracking. RADA attained feature and input level adaptation, which resulted in better adaptation without external model requirements for low-light image enhancement. Experimentation results reveal that the proposed eight models achieved SOTA results on eighteen benchmark datasets. The effectiveness of the proposed models is validated through extensive quantitative and qualitative analysis.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

$cMSE$ Corrected Mean Squared Error

$cPSNR$ Corrected Peak Signal to Noise Ratio

$cSSIM$ Corrected Structural Similarity Index Measure

F_1 F1 Score

f_n False Negative

f_p False Positive

IoU Intersection over Union

mF_1 mean F1 Score

$mIoU$ mean Intersection over Union

MSE Mean Squared Error

OA Overall Accuracy

P_r Precision

$PSNR$ Peak Signal to Noise Ratio

R_c Recall

$SSIM$ Structural Similarity Index Measure

t_n True Negative

t_p True Positive

Acronyms

ACNet Aggregated Context Network.

CD Change Detection.

CNN Convolutional Neural Network.

DA Domain Adaptation.

DL Deep Learning.

DNN Deep Neural Netwrok.

DRMNet Difference image Reconstruction enhanced Multiresolution Network.

EDCGAN Encoder Decoder based Conditional GAN.

GAN Generative Adversarial Network.

HR High Resolution.

IT Image Translation.

LLE Low light enhancement.

LR Low Resolution.

MCDNet Multi Context Dense Network.

MDAANet Masked Domain Adversarial Adaptation Network.

MM Multi modal.

MTL Multitask Learning.

OT Object Tracking.

RADA Reconstruction Assisted Domain Adaptation.

RNN Recurrent Neural Network.

RSVT Remote Sensing Vision Task.

SAR Synthetic Aperture Radar.

SCD Semantic Change Detection.

SOTA State-of-the-art.

SR Super Resolution.

SS Semantic Segmentation.

SSL Self Supervised Learning.

TMLNet Triad Multitask Learning for change detection.

UDA Unsupervised Domain Adaptation.

VHR Very High Resolution.

VT Vision task.

"Life is never fair, and perhaps it is a good thing for most of us that it is not."

~Oscar Wilde

1

Introduction

Vision task (VT) is any computational task that utilizes techniques and algorithms for the interpretation and analysis of visual inputs by a computer. Remote Sensing Vision Task (RSVT) powered with Deep Learning (DL) is revolutionizing our ability to extract meaningful insights from Earth observation data. RSVTs use image and video data captured through aerial and satellite platforms for multitemporal and multi-modal analysis in the domains of natural resource management, agriculture, disaster management, forestry, geology, hydrology, etc. The key RSVTs include image classification, SS, change identification, object detection, image fusion, Super Resolution (SR), etc. Natural resource monitoring can be done using mono-temporal analysis using Semantic Segmentation (SS) of single input instance. It can also be completed using multitemporal analysis using Change Detection (CD)

between input captured on different time instances for particular locations. Improvement in spatial and spectral details of remotely sensed inputs is required in several applications and can be achieved using Image Translation (IT). Optical remote sensing has limitations in cloudy conditions, and it can be augmented using Synthetic Aperture Radar (SAR), which has the ability to see through clouds. SAR to optical IT can produce visible interpreted images in cloudy conditions. DL techniques required significant labeled data for model training and analysis. Unsupervised Domain Adaptation (UDA) based techniques for VT adapt the knowledge learned from the source domain with ground truth to the unseen target domain with no labels. The availability of ground truth for data in a new domain is difficult and requires significant resources for generation. In Nighttime aerial tracking, data annotation is difficult due to the inability to see objects in dark conditions, but with the application of domain adaptation, it can be done.

1.1 Significance of the Vision tasks

In remote sensing image analysis, VTs are important for the interpretation of geospatial vision inputs captured for vast geographical areas. In this dissertation, we have developed deep models for four important VTs, which are Semantic Segmentation (SS), Change Detection (CD), Image Image Translation (IT), and Unsupervised Domain Adaptation (UDA). SS is a vision task that involves the classification of each pixel of inputs to predefined classes. It is used in remote sensing mapping (Land use land cover, crop mapping, natural resource mapping), monitoring (environment and infrastructure monitoring, post-classification change assessment, post-disaster analysis), and other applications. CD required two co-registered multi-temporal inputs to identify different changes between them. It is a crucial VT which is required for urban growth monitoring, deforestation assessment, natural hazard monitoring, land use change analysis, and other applications. IT is a process of converting the input of one modality (spatial, spectral) to another modality output to enable the analysis of data in different modality contexts. It has applications in SAR to Optical conversion, image fusion, image synthesis, image SR, etc. UDA is the technique to enhance the model

performance on target domain using the data of source domain. UDA is important in VTs as label creation is difficult and time-consuming for new domains. UDA techniques can be applied to any VT where source domain data is available. In this dissertation, we developed UDA techniques for SS and OT.

1.2 Motivation of the Research Work

In SS task, the field-of-view of the Convolutional Neural Network (CNN) is important but is limited by kernel size. Feature information is lost in the interpolation of features into different lower scales, which is prominent in encoder-decoder architectures. In CD task, over-fitting is very prominent in the remote sensing data due to the abundant presence of no-change class pixels, and noisy feature representations deteriorate the output performance. In IT task, temporal and multi-modal relations need to be handled effectively. In UDA based task, differences in resolution, sensors, capturing conditions, and inter-class relation between domains are common in remote seining datasets which need to be handled. Input level adaptation is difficult due to differences in the styles of domain and target images. The limitations of existing approaches for four VTs (SS, CD, IT, UDA) motivated us to undertake this research work, some of which we resolved through our proposed models.

1.3 Thesis Objectives

Motivated by the above observations, the main objectives of this dissertation are as follows:

- Develop far pixel relation based deep models for effective SS of aerial images.
- Design deep models for multitask learning based enhanced aerial CD.
- Develop multi-modal and multitemporal deep models for satellite IT.
- Propose reconstruction assisted UDA techniques for SS and nighttime aerial tracking.

1.4 Thesis Contributions

1.4.1 Deep models for aerial semantic segmentation

In the first contributory chapter, we proposed two works, namely Aggregated Context Network (ACNet) and HybridNet. In ACNet, we proposed a variant of encoder-decoder-based architecture where residual attentive skip connections are incorporated. In the second work of this chapter, we proposed HybdridNet, which extends a high-resolution network with dense connection integration to preserve the original resolution and improve parameter sharing.

1.4.1.1 *ACNet*

SS of VHR imagery is an important task where context information plays a crucial role. Adequate feature delineation is difficult due to high-class imbalance in remotely sensed data. In this work, we proposed a variant of encoder-decoder-based architecture where residual attentive skip connections are incorporated. We added a multi-context block in each of the encoder units to capture multi-scale and multi-context features and used dense connections for effective feature extraction. A comprehensive set of experiments reveal that the proposed scheme outperformed recently published work by 3% in overall accuracy and F1 score for *ISPRS Vaihingen* and *ISPRS Potsdam* benchmark datasets.

1.4.1.2 *HybridNet*

The problem of SS can be modeled as an image-to-image mapping problem where pixel-level classification is required. Pixel level classification is challenging for the high-resolution aerial image due to the presence of the tiny objects in low-frequency and more information details for such tiny objects required for dense semantic labeling. In general, encoder-decoder based architecture for SS suffers from information loss due to the up and downsampling process. To handle this, we extend a high-resolution network with dense connection integration to preserve the original resolution and better parameter sharing. We also incorporate a

lightweight self-attention module for positional attention, which results in better segmentation maps. Additionally, we use a generalized hough transform based deep voting module for pixel dependencies extraction. Experimental results reveal that the proposed model achieves the best mean intersection over union and overall accuracy in local and benchmark evaluation on the *Vaihingen and Potsdam* datasets.

1.4.2 Deep models for aerial change detection

In the second contributory chapter, we proposed two works, namely Difference image Reconstruction enhanced Multiresolution Network (DRMNet) and Triad Multitask Learning for change detection (TMLNet). DRMNet is proposed for efficient long-range dependency mapping and enhanced feature representation with multitask learning for effective CD. To further improve the performance on CD, TMLNet is proposed with triad reconstruction, enhanced backbone, multi-context local self-attention, and multitask learning.

1.4.2.1 ***DRMNet***

CD in satellite images is an important research area as it has a wide range of applications in natural resource monitoring, geo-hazard detections, urban planning, etc. Identifying physical changes on the ground and avoiding spurious changes due to other reasons like co-registration issues, change in illumination conditions, sun angle, and presence of cloud and fog is a challenging task. This work proposes a multitask learning based CD model where two parallel pipeline architectures predict change map and image difference. The proposed model takes two images and their difference as input and provides them to a Backbone Network (BN). The output of the BN is fed into the proposed Multi-Scale Attention Module (MSAM) for the effective identification of changes in multi-temporal and very high-resolution aerial images. In another parallel path, the output of the BN is down-sampled and passed to the proposed Deconvolution with a Sub-pixel Convolution Module (DSCM) to generate image difference. Two loss functions are utilized in two parallel paths to train the overall model in an end-to-end supervised setting. A comprehensive set of experiments have been carried

out, and the results reveal that the proposed DRMNet model has achieved an F1 Score improvement of 1.66% in CDD, 1.61% in SYSU, and 0.14% in LEVIR-CD datasets. It achieved an F1 score of 86.11% for the BCDD dataset with the new test image.

1.4.2.2 *TMLNet*

It is observed that an accurate CD task requires effective feature learning, which needs a noiseless representation. Moreover, CD improves with error-free image reconstruction as an auxiliary task, which needs joint feature representation by a feature extractor. In this work, we proposed a novel triad (which is a combination of input images and its difference) learning based multiresolution architecture TMLNet for effective CD. We also proposed a multi-context local self-attention module to efficiently calculate long-range pixel relations with multiple contexts. An enhanced backbone module with top-down connections and multi-scale channel and spatial attention is utilized for change map generation. It provides less noisy features extracted through the backbone. Laplacian pyramid loss is used to preserve small details in feature reconstruction. A set of comprehensive experimentations reveals that the proposed scheme achieved the SOTA result for the F1 Score, intersection over union, and overall accuracy values in seven benchmark datasets.

1.4.3 Deep models for multi-modal and multitemporal image translation

In the third contributory chapter, we proposed two works, namely Encoder Decoder based Conditional GAN (EDCGAN) and Multi Context Dense Network (MCDNet). An encoder-decoder-based model, named EDCGAN, is proposed for IT from SAR to RGB with a multi-scale attentive discriminator. MCDNet is proposed with the use of multiple spatial-space, self-attention, and dense residual attention for multi-frame SR.

1.4.3.1 *EDCGAN*

The optical images cannot reflect the actual ground information in cloudy conditions. SAR images are used to solve this for their ability to see through clouds. In this work, a novel deep learning architecture named EDCGAN is proposed. The proposed architecture is an encoder-decoder-based conditional GAN that uses multi-scale attentive discrimination to get accurate SAR to RGB IT. In addition, we have used residual connections, spatial & channel-wise attention for better feature representation. A set of extensive experimentations show that this architecture outperforms the existing SOTA method in terms of PSNR, SSIM, and FSIM_c values for the WHU-SENCity dataset.

1.4.3.2 *MCDNet*

Multi-frame SR utilizes multiple low-resolution images to generate a single high-resolution image. Multi-frame SR methods face difficulty in handling spatial and temporal dependencies of pixels. In this work, we proposed a novel architecture named Multi-context Dense Network (MCDNet) to handle spatial and temporal pixel dependencies using multiple approaches of global average pooling, multiple size kernels, and self-attention. The proposed approach improved the PSNR values by 0.29 % and 0.05 % for SR of NIR and RED bands on the benchmark PROBA-V dataset.

1.4.4 Unsupervised domain adaptive techniques for semantic segmentation and object tracking

In the fourth contributory chapter, we proposed two works, namely Masked Domain Adversarial Adaptation Network (MDAANet) and Reconstruction Assisted Domain Adaptation (RADA). MDAANet is proposed with masked domain dual adaptation, joint adversarial adaptation, domain reconstruction consistency, and feature dissimilarity alignment for domain adaptive SS. RADA is proposed with static style transfer and multi-level adaptation for domain adaptive OT.

1.4.4.1 *MDAANet*

Unsupervised Domain Adaptation (UDA) is an important task that transfers learned knowledge from the source domain to the unseen target domain. Domain shift is the major challenge faced by UDA methods, which is caused by differences in appearance, distribution, decision boundary, sensor platforms, capturing conditions, etc. The goal of UDA methods is to minimize the existing shift between domains. Domain invariant feature representation is essential for effective UDA. Mix representations combine inputs of both domains, which are helpful for the generation of domain-consistent features using self-training. Domain alignment at multiple levels is required for effective adaptive segmentation between domains. In this work, we proposed a Masked Domain Adversarial Adaptation Network (MDAANet) with a masked domain dual adaptation approach, joint adversarial alignment, consistency enforcement, and feature dissimilarity-based alignment for effective UDA. The proposed approach allows the adaptation in input, feature, and output levels, producing better UDA for SS. It increased the inter-class dissimilarities and attained input-output level combined adaptation. It achieved SOTA results on five benchmark domain adaptation tasks.

1.4.4.2 *RADA*

Visual OT is a popular research area in computer vision due to its diverse applications. Despite the impressive progress made by numerous SOTA trackers on large-scale datasets, visual OT at nighttime remains challenging because of low light (brightness) conditions, lack of contrast, very low variability among feature distributions, etc. In addition, the lack of paired (labeled) data for nighttime tracking makes it infeasible for supervised learning based modeling. UDA based tracking can resolve this issue. In this work, we proposed static image style transfer-based Reconstruction Assisted Domain Adaptation (RADA) with adversarial learning for nighttime OT. RADA attained feature and input level adaptation, which resulted in better adaptation without external model requirements for low-light image enhancement. RADA achieved SOTA results on two benchmark datasets with improvements in the range of 3.7% - 11.4%.

1.4.5 Summary of Contributions

This dissertation presented deep learning models for VTs (SS, CD, IT, UDA) utilizing remotely sensed imagery. The summary of contributions for each chapter is as follows:

- Chapter 3 contributions is summarized as:
  - In the first work on SS, a variant of the encoder-decoder-based model named ACNet is proposed for the SS of VHR images where contextual information is utilized efficiently. We proposed a dilated multi-context block in every encoder layer to capture a larger field of view.
  - In the second work on SS, we proposed HybridNet with multi-scale features, two approaches for far pixel dependence extraction, and dilated convolutions to handle multi-context enhanced field-of-view aggregated features with better parameter sharing. The hough voting and self-attention based modules are utilized for handling long-range dependency between pixels.
- Chapter 4 contributions is summarized as:
  - In the first work on CD, we have proposed a dense high-resolution connection-based DRMNet to mitigate the shortcomings mentioned above in the existing literature.
  - In the second work on CD, we proposed TMLNet, which avoids overfitting and produces precise change maps with the use of triad reconstruction, topdown enhanced high-resolution connections, local spatial and channel attention, and local self-attention.
- Chapter 5 contributions is summarized as:
  - In the first work on SAR to optical IT, we proposed an EDCGAN that utilized multi-scale attentive residual connections based generator to produce structurally rich images and attentive multi-scale full resolution discriminator with auxiliary loss for better discrimination ability.

  - In the second work of multiframe SR, MCDNet is proposed, which can effectively capture pixel relation using global average pooling, different size kernels, and dot product based self-attention

- Chapter 5 contributions is summarized as:

  - In first work on domain adaptive SS, MDAANet is proposed, which generates segmentation outputs and reconstructed inputs from the mixed masked domain inputs. MDAANet achieved input, feature, and output level alignment using the proposed Masked Domain Dual Adaptation Approach (MDDA), Joint Adversarial Alignment (JAA), Consistency Enforcement (CE), and Feature Dissimilarity based Alignment (FDA).

  - In the second work on domain adaptive nighttime OT, we proposed static style transfer based RADA for nighttime OT. Static style transfer is used to generate synthetic paired images (video frames) for supervised nighttime modeling for visual tracking.

1.5 Organization of the Thesis

The overall organization of the thesis is outlined as follows:

Chapter 1: Introduction

This Chapter presents details of VTs using earth observation data, related works, and problems of existing methods which motivates taking up the problem statements of this thesis work. The detailed problem statements of this thesis work are introduced.

Chapter 2: Background and Literature Survey

This Chapter presents the details of VTs (SS, CD, IT, and UDA) using earth observation data, followed by a discussion of recent related works.

Chapter 3: Far pixel context extraction based effective semantic segmentation

In this Chapter, two deep models (ACNet, HybridNet) are proposed which can effectively capture far pixels relations for aerial SS.

Chapter 4: Multitask learning enhanced aerial change detection

This Chapter presented details of two proposed deep models (DRMNet, TMLNet) for aerial CD, which utilize multitask learning to avoid prominent overfitting.

Chapter 5: Multimodal and multitemporal models for image translation

In this Chapter, two IT models for multi-modal (EDCGAN) and multitemporal (MCD-Net) inputs are proposed, which utilize multiscale discriminator attention and multi-context features for IT.

Chapter 6: Reconstruction assisted unsupervised domain adaptive techniques

This Chapter presented details of domain adaptive schemes for SS (MDAANet) and nighttime OT (RADA). Both approaches used reconstruction as the secondary task, which contributed to better domain invariant feature representation.

Chapter 7: Conclusion and Future Perspectives

This chapter presented the conclusion of this thesis works with an outline of possible future extensions of current work and the way forward.

"Everything is hard before it is easy."

~Johann Wolfgang von Goethe

2

Background and Literature Survey

Deep Learning based approaches provide the ability of automatic feature engineering, which is essential in vision applications of the remote seeing domain. In this work, we take up four important Remote Sensing Vision Task (RSVT)s which are Semantic Segmentation (SS), Change Detection (CD), Image Translation (IT), and Unsupervised Domain Adaptation (UDA). We proposed novel deep learning models for these tasks to mitigate the shortcomings of existing approaches. Our proposed approaches have achieved state of the art results on several benchmark datasets for these four important vision tasks of remote sensing. The details of these tasks are as follows:

2.1 Semantic Segmentation

Semantic Segmentation (SS) of the aerial image is an essential research topic because of its usage in different satellite image-based applications such as land-use mapping, progress monitoring, building footprint extraction, water body extraction, etc. Semantic segmentation is an image-to-image mapping task. Here, the main objective is to classify each pixel of input into predefined semantic classes. The visualization of this task is represented in Fig 2.1. Traditional shallower machine learning approaches such as Random Forest (RF), Support Vector Machine (SVM), on conventional features like Scale-invariant feature transform (SIFT), Histogram of oriented gradients (HOG) utilized predefined features for SS. In recent times, Deep Neural Netwrok (DNN) based SS models achieved improved performance over handcrafted feature-based approaches. Thus, the focus of the research has shifted from feature engineering to network modeling. Among DNN based SS models, encoder-

Figure 2.1: *Visualization of SS task on DSTL [1] dataset.*

decoder-based architectures [4–11] are popular due to their simplicity and extendability. The preservation of the original resolution of input [12–15], is also frequently reported. The major disadvantage of these approaches is that they are primarily computational intensive and require higher graphics memory. It is also observed that the field of view plays a vital role in SS. Thus the large kernel size is a natural choice to increase the effective field of view of the filter [16]. Alternatively, few approaches use atrous convolution [7, 11, 17–19], which utilizes a hole in between the kernels to result in a large kernel effect. The large

kernel achieves a large field of view without increasing resource requirements. Attention is a mechanism to highlight the important features and utilized in recent work [18, 20–22]. Dense connection based neural networks [19, 23] provide parameter sharing and produce better results for SS. A conditional random field is used by some approaches [24, 25] as a post processing for enhancement of the segmentation results. Super pixels with CNN based approach [26] is also proposed for effective SS. Some recent approaches use the generalized hough transform to find distant pixel interaction [27–31].

2.2 Change Detection

The Change Detection (CD) task for remote sensing images has many applications in urban growth monitoring, disaster analysis, agriculture applications, climate change, forest loss, etc. In Fig. 2.2, the visualization of the change detection task is shown. It can be modeled as the measurement of pixel and object level information change in bi-temporal remotely sensed images of a particular region. The CD between input pairs can be mathematically represented as

$$CM = f(I_1, I_2) \tag{2.1}$$

where f is a mapping function that takes two different timestamp inputs. The images I_1 and I_2 are the inputs, and CM is the produced change map. Traditional handcrafted features-based approaches use Clustering [32], Change Vectors [33], and other methods for estimating the mapping function f. In algebra-based approaches [33, 34], this function is assumed to be an algebraic operation computed on input images. Transformation based methods [35, 36] performed different modifications to the inputs, and the change map is calculated on transformed inputs. The object-based approaches [37, 38] first extract the object-related information and then generates change maps. This function is calculated using neural networks in deep learning based CD. The difference in features approximately represented the changes, and some works utilized the feature differences [39–48] in the neural network. Further, this approach is expanded for the use of multi-scale feature differences [43, 44, 49].

Focusing on the relevant area of the input is important as it filters out the irrelevant regions. The attention-based [39–42, 48–57] methods do this using attentions maps generated on local and global levels. The calculation of global attention is computationally intensive, so a resource-efficient approach [47] is required for global pixel interaction. Scale plays an important role in the identification of changes. Few methods [40, 42, 45] calculated change maps at multiple scales. Adversarial models are also used in several new approaches [52, 58–60]. Few methods [39, 56, 61] are based on increasing kernel size. Dilated convolutions based networks are proposed to take advantage of multiple context features. Super-resolution [62] based CD is proposed where multiple resolution inputs are generated. CD tasks used multi-temporal inputs. The Recurrent Neural Network (RNN) [46, 63] and structural difference [45] based networks are proposed to utilize the temporal relation between inputs. Feature sharing between the inputs is essential, which are effectively extracted by siamese-based networks [40, 43, 46, 64]. Multitasking-based methods [48, 51] are proposed where additional tasks are learned with the CD backbone. Metric learning based method [53, 54, 64–66] is proposed to handle class imbalance problem. Autoencoder based approaches [67–69] are also utilized for homogeneous and heterogeneous CD.

Figure 2.2: *Visualization of CD task. Here, the white color is the actual change detected by the model depicted in the right image, and the left and middle images are the inputs.*

2.3 Image Translation

In cloudy conditions, an optical sensor can not capture the ground reality, but can be captured using a SAR sensor. The visualization of SAR images is difficult due to the presence of various noises. SAR to optical image translation is a task that focuses on converting SAR images to optical images for better visualization, as shown in Fig. 2.3. High Resolution (HR) satellite images are not easily available. Low Resolution (LR) satellites mostly have higher temporal resolution than HR satellites. It makes satellite super-resolution a task important as it makes super-resolved HR data available.

Figure 2.3: *Visualization of SAR to Optical IT task. Here, left image is SAR input, middle image is sentinel 2 output, and right image is predicted output.*

In Fig. 2.4, multi-frame SR task is visualized. Non-deep learning based approaches for super-resolving satellite images, in general, utilized edge preservation [70], wavelet transform [71, 72], spline interpolation [73], sparse coding [74] and genetic algorithm [75]. In deep learning based methods, adversarial learning [76–81], RNN [82], 3D convolutions [2, 83–87], transformer [88, 89] and uncertainity estimation based approach [90] are proposed. New datasets for satellite image SR are introduced in some works [91–95]. Utilization of multi context features for effective multi-frame satellite SR is missing in the present approaches.

2.4 Unsupervised Domain Adaptation

Domain Adaptation (DA) is an emerging area of research in the remote sensing field. Ground truth creation is an extremely tedious task. SS requires pixel-level details which makes an-

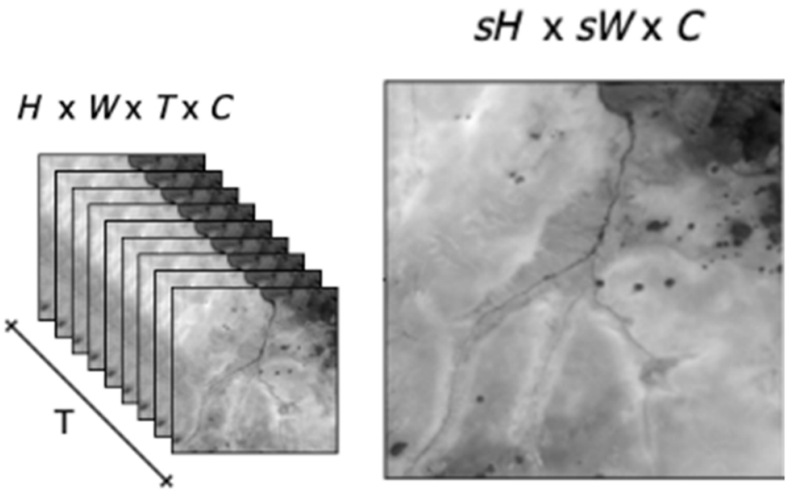


Figure 2.4: *Visualization of multi-frame SR task [2]. Here, the left image is the input low resolution image, and the right image is the predicted HR output.*

notation creation more difficult. DA provides the ability to adapt the model to other unseen domains which makes it an important task with many applications. In Fig. 2.5, the DA task is represented. Unsupervised Domain Adaptation (UDA) is a type of DA where target domain labels are not available during training. UDA task can be implemented with alignment at different levels, which are input, feature, and output. Multi source methods [96–98] used multimodal inputs for domain alignment. Self-training is an approach to generate pseudo labels of target domain during the training with source domain data, and it is utilized by several methods [99–107]. Consistancy based approach [107–111] enforces consistency constraints on outputs of different augmented inputs. Nonadverserial methods have been used by some methods [112, 113].

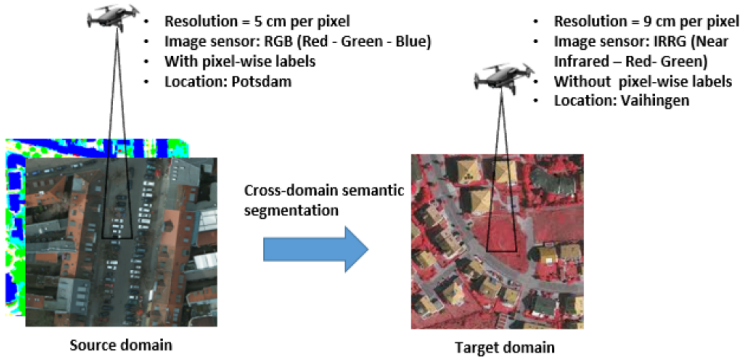


Figure 2.5: *Visualization of DA task for SS [3]*

Correlation between features is proposed in [100,114] for effective alignment. KL divergence-

based [106, 115] approach between outputs and features are used for adaptation. Output level adaptation is proposed in [106, 116–119] which aligns the category outputs of domains to reduce domain shift. Class centroids are calculated for output level adaptation in [106]. The feature level alignemnt is utilized in [100, 106, 118, 120] to attain domain invariant features.

Input level alignment tries to match the visual and statistical characteristics of both domains. It is implemented in [110, 121–126, 126–128] with use of scale feature [128], IT [110, 119, 124, 125], and style transfer [107, 119]. Dual Generative Adversarial Network (GAN) based methods used two GAN for IT [124, 125]. All level alignment is proposed in [98] for adaptive segmentation. Cycle GAN based approach [124] is used to translate the inputs of one domain to another to achieve input level alignment. Memory construct [129] is proposed to store the invariant representation of the complete dataset inputs. Curriculum based approaches [130] used easy to hard learning for DA. Calculation of entropy values of both domains and their alignment is used in some methods [112, 130, 131, 131–133]. Adversarial based approaches [96, 98, 102, 103, 106, 109, 115, 117, 118, 124, 127, 131, 133–141] relies on maximization of confusion between domain representations. Student teacher [102, 111, 126] based methods minimized noisy labels with the use of a separate teacher network. Attention methods [103, 106, 129, 133, 137, 139] is used to highlight the common representations of both domains. Covariance measures the relationship between the domains and is used to align the representation [103, 104].

2.5 Literature Survey

2.5.1 Semantic Segmentation

2.5.1.1 Semantic segmentation and Hough voting

There are plenty of SS algorithms that have been proposed in the recent literature. Long et al. proposed a fully convolutional network-based SS [4]. It replaced fully connected layers, which helps in the better preservation of spatial features. Onneberger et al. [5] introduced

encoder-decoder based architecture with skip connections which allowed utilization of low and high-level features. Yu et al. [7] proposed dilated residual network (DRN) that utilized dilated convolution to increase the receptive field of the kernel. The original resolution of features contains detailed information. Several works are being proposed to utilize feature information at original resolution. Pohlen et al. [13] introduced full-resolution residual networks that used residual stream to operate on full resolution and pooling stream process multi-resolution input. These two streams are fused using a full-resolution residual unit (FRRU). Fourure et al. [12] proposed GridNet that used multiple interconnected multi-resolution paths. Sun et al. [14] and Wang et al. [15] proposed High-Resolution Network (HRNet) for vision problems. It preserved the HR information for precise and semantically strong output feature map generation.

Attention is a mechanism to focus on essential features that need to be focused on spatial and channel dimensions. Recent work exploits this utility for better delineation of objects in feature maps. Fu et al. [18] introduced a dual attention network that used dilated FCN with two attention modules for spatial and channel dependencies. Yue et al. [20] proposed a compact generalized non-local network (CGNL) for extraction of long-range spatial and temporal dependencies. It learned the correlation between two positions across channel dimensions with a compact representation of kernel functions with efficient Taylor expansion. Li et al. [21] proposed a local distribution module for adaptive distribution of features extracted from the global aggregation module to delineate small patterns effectively.

Hough transform is used to identify the object with specific shapes. Samet et al. [27] proposed HoughNet that utilized log-polar voting module for near long-range class conditional evidence. Milletari et al. [28] proposed Hough-CNN that utilized the hough voting method. Novotny et al. [29] introduced semi-convolutional networks and showed the relation of it with hough voting. Qi et al. [30] proposed VoteNet that used end-to-end differential hough voting. Sheshkus et al. [31] proposed a fast hough transform layer to find vertical and horizontal vanishing points using two branch networks. Hough voting based methods suffer from false positive detection due to the impact of texture region and are generally slow [142].

In deep voting based methods that automatically extract texture and other features, voting space increases with the use of a large number of voting regions [143]. The voting module can be added with a deep network during the training phase [27] or can be used at inference time [30, 31]. Deep voting based methods allowed the extraction of local and global pixel interaction. We used hough voting in the network as ended-to-end learning to take advantage of backpropagation and utilized the generated heatmaps for better representation of features.

2.5.1.2 Semantic segmentation of aerial imagery

Liu et al. [24] proposed high order conditional random field with features extracted using a pretrained fully convolutional network for multi-sensor SS. They used multinomial logistic regression for feature extraction from LiDAR input. Liu et al. [144] presented ScasNet to utilize multi-context information and process it in a self cascading manner. Marcos et al. [145] utilize rotation equivariance information and encode it in the network. Yue et al. [146] proposed a Tree-CNN block, and each node of it was generated based on confusion matrix and tree cutting algorithm. Sun et al. [8] pointed out the problems of structural stereotype and insufficient learning faced by encoder-decoder architecture. They proposed ResegNet with random sampling training and ensemble inference to handle it. Marmanis et al. [147] integrated edge detection in SS pipeline with the inclusion of class boundary network at the start. Nicolas et al. [148] proposed V-FuseNet, which takes early fused input of input bands and DSM information. Audebert et al. [16] used a multi-kernel convolutional layer to utilize multi-context information using different filter sizes. Maggiori et al. [149] utilized MLP network to combine features of different resolutions and combine local and global information. Volpi et al. [9] used an encoder-decoder network with a full patched learned upsampling network. Liu et al. [10] proposed an Hourglass shape network that utilized inception module in the network, residual skip connection, and postprocessing module based on belief propagation. Diakogiannis et al. [11] proposed ResUNet-a, which is based on an encoder-decoder architecture with residual connection, dilated convolution, and pyramid

pooling. It also utilized the tanimoto loss function with complement. Paisitkriangkrai et al. [25] combination of convolutional neural network classifier and random forest classifier using handcrafted features. Conditional random field used for postprocessing of extracted features. Chen et al. [23] introduced SNFCN and SDFCN network frameworks that utilized dense shortcut connections. Overlay strategy is used for postprocessing of output. Sherrah et al. [150] used a fully convolutional network with no down-sampling and hybrid FCN, which combined pretrained features with DSM features. Liu et. al. [19] proposed DDCM with densely dilated connections. Liu et al. [22] proposed a self-attention and edge enhancement network that utilized spatial self-attention and edge features extracted using laplace operation.

2.5.2 Change Detection

The original resolution of the input contains rich feature information, and feature extracted at the original resolution streams are used to avoid information loss. UNet++/ Nested-Unet [151, 152] is a variant of encoder-decoder architecture that utilizes feature information at the original resolution. Fang et al. [50] proposed SNUNet, in which siamese networks with Nested-Unet and combined attention module are used for different features. It produced a semantically rich change map that utilized shallow and deep layer features. Zhang et al. [42] proposed DiffUNet++, which combined differential pyramid with UNet++. The dense skip connections between same-level resolutions are used in UNet++, with a difference of input pairs as input generated by differential pyramids utilized. It resolved the disparity of changes at multiple scales. Peng et al. [41] presented a difference-enhancement dense-attention convolutional neural network (DDCN) with spatial and channel attention in upsampling units to focus on spatial context and features of different levels. For intuitive change map generation, a DE unit is proposed that transforms feature difference maps into new feature space. Wang et al. [40] proposed an attention-based ADS-Net that utilized an adaptive attention fusion module. Features are extracted using a shared encoder and multi-scale input pairs features. Its difference is fused and utilized by channel-wise and spatial attention

modules at different scales for enhanced change map generation.

Difference map at multiple scales helps in better change map generation. Zhang et al. [43] proposed a feature difference convolutional neural network (FDCNN) that generates a feature difference map at different scales and feature depths using pretrained network weight. Feature fusion network is used to learn the trainable weights during training. It avoided over-fitting and required a few training samples. Jiang et al. [44] proposed a dual pathway CD network (DP-CD-Net) with ResNet pretrained weight for feature extraction. Multi-scale feature difference maps are generated using different layers of encoder and decoder separately. It improved the existing feature difference-based network with the proposed use of a multi-scale difference map after the decoder to take advantage of the HR information of the encoder with semantic-rich decoder features. Shi et al. [49] proposed a deeply supervised attention metrics based network (DSAMNet) that utilized multi-scale feature differences with channel and spatial attention. They introduced the SYSU CD dataset for complex change types.

Bai et al. [46] proposed edge guided recurrent convolutional neural network (EGRCNN) that utilized prior knowledge of edges and discriminating features for building CD. LSTM based module is proposed to extract and enhance the difference information. Hou et al. suggested a HR triad network (HRTNet) [45] to take advantage of temporal information using a difference map as one of the inputs. It is produced using pixel and textual details of the input pairs. Dynamic convolutions based inception module is used for efficient operations and to recognize multi-scale changes.

Multi-context based approaches are used to provide a larger field of view. Song et al. [56] proposed an attention-guided end-to-end CD network (AGCDetNet) with an attention module that utilized prior knowledge in self-attention. Channel attention with atrous convolutions is used for multi-context attentive representation. Li et al. [61] used a multi-scale fully convolutional based network (MFCN) with a multi-scale convolution module. It consists of a multiple-size kernel for multi-context feature extraction. They used cross entropy with a dice loss function to handle a highly unbalanced dataset. Ke et al. [39] intro-

duced a multi-level change contextual refinement network (MCCRNet) that utilized change contextual representation generated using multiple self-attention modules for the generation of a better change map. Chen et al. [47] introduced efficient transformer based CD using the visual token for context mapping to avoid the high computational requirement of self-attention based methods. The enhanced pixel features of individual input images through token-based context mapping are used to calculate feature differences and final change map creation. Liu et al. [51] proposed a dual task constrained deep siamese convolutional network (DTCDSCN) that performed building CD and building extraction simultaneously with a common feature extraction backbone. Dual attention module to utilized to further enhancement of feature representations.

Chen et al. [53] proposed spatial-temporal attention neural network that used self-attention maps. A basic attention module is proposed for global feature dependencies, and multi-scale spatial-temporal dependencies are extracted using a pyramidal attention module (PAM). Deep metric learning based contrastive loss is used. They also introduced the LEVIR-CD benchmark dataset for building CD. Chen et al. [54] introduced a dual attentive fully convolutional siamese network (DASNet) that used deep metric learning for feature distance calculation and proposed weighted double margin contrastive loss to mitigate class imbalance problems. Du et al. [66] proposed a bilateral semantic fusion siamese network (BSFNet) that utilized deep and shallow networks with channel-wise and spatial attention for extracting features. These deep and shallow features are combined using a semantic fusion module with the use of an antiactivation function. A scale-invariant sample loss is also proposed for metric learning based CD. Xu et al. [64] proposed MFPNet that used multi-directional and adaptive weighted fusion. It allowed the information to pass through multiple paths. A perceptual similarity module is proposed to utilize perceptual loss to eliminate the disadvantages of cross-entropy loss functions.

Liu et al. [67] proposed a deep coupling network for heterogeneous CD. Features are extracted using a symmetric network of convolutional and coupling layers, and a change map is produced using a threshold based difference map. Kalinicheva et al. [68] utilized

autoencoders to transform input to feature space, and then the difference map was calculated to find changes. Luca et al. [69] used a convolutional autoencoder (CAE) to extract hierarchical features. These features are used to generate a change map with a hierarchical loss function. Kalinicheva et al. [153] proposed temporal evolution graphs to cluster types of changes from multiple change segmentation maps produced using autoencoder with gated rectified units (GRUs).

Chen at. al. [58] proposed instance-level change augmentation (IAug) using adversarial learning and image blending to increase the number of change samples in the dataset. A deep siamese-based convolutional network is used to utilize the enhanced dataset, and change maps are produced. Zhao et al. [52] proposed an attention gate generative adversarial adaption network (AG-GAAN), which created a change map using a generator. Those are validated by ground truth using a discriminator network with domain similarity loss function. Hou et al. [59] utilized WNet for feature extraction and change map generation. It extended the CD problem as an adversarial learning problem, and conditional GAN with WNet as a generator is used. Wu et al. [154] utilized superpixels of input pairs using the SLIC algorithm. The change vector is computed for all superpixels, and the final change map is produced. Zhang et al. [155] proposed end-to-end superpixels enhanced CD network (ESCNet) with two superpixel sampling networks for features extraction. Extracted features are used by UNet based CD network with adaptive superpixel merging.

2.5.3 Image Translation

2.5.3.1 Multi modal Translation

In the past decades, many algorithms and methods have been proposed to make Synthetic Aperture Radar (SAR) images look better for visualization. For more suitability, Jiao et al. [156] proposed a SAR image visualization algorithm to map the high dynamic range SAR amplitude values to low dynamic range displays via reflectivity distortion preserved entropy maximization. For the further recognition of ocean vessels and aircraft, an adaptive two-scale enhancement [157] method was proposed. Single polarized SAR images can

be improved by encoding its pixels with various colours using the pseudocolour encoding algorithms [158], [159], [160], which can be used in other SAR image applications.

Deep learning has emerged very rapidly in the areas of computer vision to target various issues like Image-to-Image translations and identifying Flood maps from SAR images [161]. GANs [162] have performed exceptionally well in recent years for performing Image-to-Image translations. They are fast, reliable, and accurate. For satellite imagery, the neural network is used for super-resolution, ship detection, land cover classification etc. A GAN is used to create pictures from arbitrary noise. It consists of the generator and discriminator. The generator tries to understand the underlying distribution of the data and predicts the output image from the random noise. The discriminator is being tried to be fooled by providing generated images and true images. The discriminator has to identify whether the supplied image is true or fake. It progresses as a minimax two-player game. Various GANs are proposed for image-to-image translation. Due to uncontrollable translated results in the target domains, conditional GAN [163] helps in controlling the target domain by supplying additional Label to the generator and helps in training the GAN in a specific domain.

Due to the different ground sampling distances (GSD) of the Sentinel 1 and Sentinel 2 satellites, there is always some spatial inconsistency between the input image and the target image. Traditional GANs may not be able to train properly over such data. Hence CycleGAN [164] has been introduced to consider cyclic consistency between the input and target image. These GANs have performed well in translating ideas with large spatial inconsistency. But they are unable to preserve land-cover information properly in case of SAR to RGB Satellite IT. Pix2Pix [165] has been a breakthrough for the image to image translation and has performed well for Satellite IT as well. Pix2Pix can well characterize land-cover information, but it produces a blurry effect on complex structures. The CycleGANs are improved by combining their property with the property of Pix2Pix [165]. A supervised Cycle-Consistent Adversarial Network (S-CycleGAN) [166] has been the state-of-the-art method for Satellite IT so far. The generator and discriminator of S-CycleGAN have the same network architecture as the Pix2Pix which is U-Net and PatchGAN, respectively. The

loss function of Pix2Pix is modified to consider the cyclic-consistency loss, which computes the loss at the pixel level. Hence, the input SAR image and RGB image should be paired and there should be minimum spatial and temporal inconsistencies between the images. It is able to capture land-cover information and also able to preserve structure information. Nevertheless, it doesn't work for all types of land structures.

Zhang et al. [167] proposed the use of intensity-hue-saturation and wavelet-based fusion, resulting in better spatial details translation of SAR images. Niu et al. [168] used conditional GAN for SAR to RGB translation and validated the generated result using land cover classification. Bermudez et al. [169] proposed the use of multitemporal images as a condition variable for generating better images using conditional GAN. Zhang et al. [170] introduced IT for session varying multitemporal image CD task using cycle GAN with style discriminator for the sessions. Enomoto et al. [171] also used conditional GAN for SAR to RGB translation. Zhang et al. [172] conducted a comparative study of the effect of edge information, and polarization details of SAR images on the conditional GAN based optical translation for various landuse cases. Zhang et al. [173] used multi-scale feature guidance for optical translation and use of discrete cosine transform to remove the high frequency noise of the SAR images. Zhang et al. [174] used combination of supervised and unsupervised models using image reconstruction from distorted input. Christovam et al. [175] use the conditional GAN based optical translation task for cloud removal from the crop dataset. Yu et al. [176] proposed the use of attention mechanism in conditional GAN with discrete cosine transform based loss function. Guo et al. [177] proposed edge preserving convolution to produce structurally rich images using adverserial learning. Zhang et al. [178] used a neural partial differential equation based model for explainable translation with finer details in the generated image. Fu et al. et al. [179] utilized a residual network with hybrid loss for reciprocal IT. Zhao et al. [180] utilized variation GAN and introduced a new SAR to RGB translation dataset named SARGB. Yang et al. [181] proposed enhanced conditional GAN with parallel feature fusion for contour information preservation, multiple size receptive fields, and the use of chromatic aberration loss. Doi et al. [182] used regional details

extracted through a pretrained convolutional neural network as auxiliary feature information for the generator of conditional GAN. Ji et al. [183] focused on unpaired multi-domain optical translation using cycle consistency loss in the adversarial network. Wang et al. [184] used a combination of despeckling network and colorization network for the translation task. These GANs have shown very good results for SAR to RGB Satellite IT. But they all are unable to capture finer details of the image, and hence there is a scope for improvement in this area due to less PSNR of the image produced by the state-of-the-art method.

2.5.3.2 Image Translation (Super Resolution)

Farsiu et al. [70] proposed the use of L_1 norm minimization as regularization for multi-frame reconstruction based SR through edge preservation and denoising. Demirel et al. [71] proposed a dual tree based wavelet transform approach that used high-frequency sub-band image and original input. The results are further improved with the use of discrete wavelet transform to generate four sub-band images and high-frequency sub-bands with low resolution input and difference image of low-frequency sub-band, and input is used [72]. Kawulok et al. [75] used evolving imaging model that utilized genetic algorithms with evolving kernels to find an imaging model. Anger et al. [73] used a two-stage approach of image registration and fusion where limitations of trigonometric polynomial based interpolation in fusion step are resolved with the use of high-order spline interpolations. Hakim et al. [185] proposed a variational SR framework with automatic weighting parameters of L1 and L2 norms that can handle impulse and gaussian noise. A combination of total variation and total variation of first derivatives is used as regularization to achieve the generalized solution. Kato et al. [74] proposed double sparsity to handle image registration and sparse coding as a single objective for multi-frame SR, which is faster than other approaches.

3D convolution provides an effective way to extract temporal relations. In 3D convolution based approaches, Molini et al. [83] proposed a combined strategy of single-image super-resolution and feature fusion. The input LR images are upsampled using bicubic operation and passed to 2D convolution based network for super-resolution. Later all enhanced images

are fused using 3D convolutions based fusion module. They further enhanced the network [84] with the inclusion of graph convolution layer with capturing of nonlocal features. Dorr et al. [85] used wide activation neural network that used a 3D convolution based layer followed by a pixel-shuffle layer for super-resolution of the inputs, which are aggregated with 2D convolution network based features produced on the input of all frame mean value. Salvetti et al. [2] introduced a residual attention model that used a 3D convolution based neural network for learning the spatiotemporal relationship between the LR features. Temporal attention using global pooling operation on the temporal dimension is also utilized. Ibrahim et al. [86] proposed a 3D convolution based network with residual and dense connections.

In adversarial learning based methods, Jian et al. [76] proposed an edge enhancement network that used GAN for HR image generation, which is further improved using a fusion of edge and texture extracted using separate networks. It solved the problems of poor perceptual image generation and noisy artifacts using edge masks for the removal of noisy regions. Ma et al. [78] proposed saliency discriminated GAN to avoid pseudo-texture generated in adversarial methods. The generator utilized the weakly supervised saliency maps, and paired discriminators were used for salient and non-salient regions to produce more realistic super resolved images. Mostafa et al. [79] used a squeeze-and-excitation module in the generator to improve the perceptual quality of super resolved image. Wasserstein distance is used in training to handle the vanishing gradient problem. Shin et al. [77] proposed a perceptual based method to generate object shapes and scale space filtering using a shape discriminator. It emphasized the object shape details in generated output and avoided artifacts. Wang et al. [80] attention constructs in the generator for enhanced spatial information extraction and discriminator network to improve discriminative ability. Wang et al. [94] proposed a conditional GAN based approach and introduced a multi-sensor satellite super-resolution dataset named OLI2MSI. Razzak et al. [95] curated a new multi-temporal super-resolution dataset for RGB satellite images.

Temporal relations can be extracted using Recurrent Neural Network (RNN) and transformer-based methods. Arefin et al. [82] utilized a RNN for the fusion of encoder features from

LR inputs. The deconvolution layer based decoder is used for the final reconstruction of the super-resolved image. An et al. [88] introduced the use of a transformer module for multi-frame satellite super-resolution. Residual based encoder module, transformer based fusion module, and pixel-shuffle layer based decoder module are used. Xiao et al. [89] used a spatial-temporal transformer based neural network to enhance the spatial-temporal resolution of satellite video.

Zhang et al. [186] introduced two streams of convolutional neural networks to utilize the spatiotemporal features. It extracts the temporal relation between multiple LR images and uses it for single HR image generation. Kawulok et al. [187] combined the deep learning based SR for single input with genetic algorithm based evolving imaging models. Residual neural network is used for single image SR, and super resolved images are fused using evolving imaging model. Aburaed et al. [188] used multi-scale wavelet decomposition with a convolutional neural network for approximation of high-frequency details. The input image is transformed to YCBCR color space, and the luminance component is precisely super resolved using a neural network because human eyes are more responsive to this component. Deudon et al. [189] proposed the use of recursively fused features extracted through an encoder using multiple LR images. ShiftNet is proposed to register and align the predicted output with the ground truth. Yan et al. [190] proposed a multi-task learning-based multiple-blur-kernel base framework that can handle low resolution input with unknown degradation function. Gaussian kernel generator produces random isotropic and anisotropic Gaussian blur kernels. Valsesia et al. [90] proposed PiUNet that utilized temporal permutation invariance property to improve the super-resolution performance. A method for aleatoric uncertainty quantification is also presented and correlated with the temporal permutation of LR images.

Daniel et al. [191] introduced the variation loss function to generate perceptually rich super-resolved output with enhanced edge and texture features. Lee et al. [192] introduced the use of quality maps present for LR inputs in the calculation of temporal attention maps. Liu et al. [193] used a multi-scale deformable convolution neural network to produce HR

arctic sea images. Sun et al. [194] proposed an adaptive nonlocal mean algorithm for super-resolution of thermal images with better geometric features reconstruction. The proposed algorithm is extended for GPU acceleration for faster computing.

Few thermal and hyperspectral satellite image and video SR methods have been proposed. Mandanici et al. [195] introduced a multi-frame algorithm for super-resolution of LR thermal images with minor differences in the viewpoint. Mei et al. [87] used 3D convolution-based neural network for spatial and spectral super-resolution using multispectral and hyperspectral satellite images. Liu et al. [196] introduced joint learning of super-resolved frames and blur kernel prediction for satellite video super-resolution. Yi et al. [81] proposed GAN based approach for satellite video super-resolution. Temporal group projection is used for the fusion of temporal details with multi-scale residual and deformable convolution based generator.

Martens et al. [197] introduced the PROBA-V dataset with cloud coverage details for multi-frame satellite super-resolution. Baseline results are provided on this dataset with a multi-layer convolutional neural network. Kawulok et al. [92] extended multi-frame super-resolution for Sentinel-2 data using deep learning methods and introduced a super-resolution dataset. Nguyen et al. [93] used a self-supervised based network with shift and add method to combine features. A synthetic dataset for Sentinel-2 L1B products is also introduced.

2.5.4 Unsupervised Domain Adaptation

2.5.4.1 Unsupervised Domain Adaptation for semantic segmentation

Domain shift is a significant problem faced by adaptation methods. Yan et al. [134] proposed adaptive DA (ADA), which utilized a domain similarity discriminator to reduce the shift in the domains. Tasar et al. [198] pointed out that the shift in spectral distribution is significant in different domains and proposed ColorMapGAN that aligns the source domain spectral features with the target domain without changing the semantic information of the source input. Benjdire et al. [3] utilized unpaired IT between source and target domain using the generative adversarial network. It helps in feature alignment between two domains. Cai

et al. [110] proposed BiFDANet, which used bidirectional IT to minimize domain shift and semantic consistency loss. Scale information is vital in DA of different resolution satellite imagery. Deng et al. [128] proposed the utilization of an additional scale-aware discriminator to handle scale information effectively. They also proposed a scale attention module for the segmentation network.

Chen et al. [112] proposed a non-adversarial network to learn class-aware distribution differences as regularization. Entropy minimization at the target domain is utilized to generate high-probability pseudo labels. Deng et al. [116] utilized output space alignment for DA using adversarial domain learning, which reduced the domain shift. Chen et al. [118] used a class-level local and global domain discriminator. The local discriminator helps in feature-level class-specific alignment, and class-level alignment is achieved using the output domain discriminator.

All level alignment is better for domain adaptive SS tasks. Peng [199] proposed a whole-level DA network with a Wallis filter for input level alignment, adversarial learning to reduce feature level domain shift, and mean teacher network for output level alignment. Ji et al. [98] proposed a full-space DA network (FSDAN) to match the domain in input, feature, and output levels. It used multiple generative adversarial networks and perceptual loss. Toldo et al. [113] proposed a non-adversarial approach for latent space alignment between domains. It extended feature clustering with orthogonality and sparsity constraints for SS tasks. Orthogonality constraint minimized the inter-class relation, and sparsity ensured the minimum active features. Zhang et al. [130] proposed curriculum learning that adapts segmentation networks based on uncertainty ranking of target domain patches. It utilized semantic and entropy-based domain classifiers for local and global features.

Guo et al. [137] introduced a selection of class-relevant features using the attention mechanism, which is utilized for group-wise feature alignment using adversarial learning. Spatial attention emphasizes relevant features of both domain classes and reduces the domain shift. Zhao [109] utilized geometric-consistency constraint to maintain the semantic information of inputs. Joint training with translated source images with geometric constraints is per-

formed to improve the segmentation performance. Chen et al. [117] proposed the utilization of class-level domain alignment with class attention and adversarial adaptation at the domain and class level to avoid negative transfer. Iqbal [200] proposed weakly supervised DA for built-up areas in satellite and aerial images. It used weak labels available to denote objects' presence, which improved the adaptation. Yan et al. [140] proposed the utilization of source and target domain features in adversarial learning. It used a discriminator to distinguish the features of input pairs from the same and different domains, which helps to extract domain invariant features due to the simultaneous use of source target details. Zhang et al. [114] proposed subspace alignment to handle interclass similarity. It used correlation maximization to avoid irrelevant mapping between domains.

Lin et al. [139] proposed Xnet, which used dual adversarial adaption with separate domain generators and task-oriented attention for remote sensing image classification. The separate domain generators handled the large domain gap and adversarial adaptation between generated features of source and target domains and between target and common domain features, resulting in better alignment. Gao et al. [124] utilized a dual GAN based approach that used cycle-consistent style transfer for input level alignment of domains. The transformation consistency with self-supervised learning on the target domain is also added to enhance the invariant feature extraction. Zhao et al. [125] also proposed DualGAN based approach named ResiDualGAN, which performed unpaired IT with SS to achieve domain adaptive segmentation. Liu et al. [115] proposed an adversarial DA framework (KL-ADDA) that used KL distance between the features of the discriminator for the source and target domain.

Liu et al. [135] proposed the conversion of features extracted by the deep network to two-dimensional feature curves. The conditional generative adversarial network is employed to reduce the dissimilarity between source and target two-dimensional curves. Yan et al. [111] proposed a cross mean teacher based approach to resolve the missing pixels problem of boundary regions after classwise alignment. Two pairs of teacher-student networks are used for cross-consistency and better pseudo-label generation. Liu et al. [105] used a feature

pyramid network (FPN) with self-training that used weighted pseudo labels. The FPN with pyramid pooling module allowed better segmentation maps, and self-training enabled the domain alignment. Zhao et al. [102] proposed ST-DASegNet that used disentangled adaptation with separate source and target domain student and teacher networks. Adversarial learning at the feature level is utilized for domain alignment with an exponential moving average to update the weight of the teacher model.

Wang et al. [106] proposed adversarial learning based method that utilized global alignment, which is further improved by local alignment using attention maps. Output level alignment is achieved using class centroid with student-teacher divergence loss. It minimizes the intra-class variation and increases the inter-class variance. Chen et al. [107] proposed consistency regularization to handle cross-domain and geometric shifts. Cross DA is achieved using Fourier transforms and ClassMix adaptation. Geometric shift is handled using a simulation module for view distortion and scale variation. Ni et al. [119] proposed integer programming based output alignment for DA, which used class relation constraints. It minimized the inter-class variation and increased the intra-class similarity. Zhang et al. [120] utilized feature-level relations using similarity losses. Strong and weak similarities in the information of nearby outputs with source features are used during alignment.

2.5.4.2 Unsupervised Domain Adaptation for nighttime aerial tracking

The correlation filter based object trackers use a template of the target object to find the most similar regions in the subsequent frames. They often require special handcrafted feature extraction techniques for processing. Galoogahi et al. [201] used background based negative samples extracted through handcrafted features in correlation filters for efficient Object Tracking (OT). Li et al. [202] proposed varying spatiotemporal regularization, which utilized global and local response variations details in the object tracker. The Siamese network-based approaches focus on finding a generic similarity function. They generally use a pair of identical common weighted neural networks to find the correlation between the target object and the search region and take the benefit of not requiring any unique feature

engineering due to the end-to-end learning possible. Chen et al. [203] proposed a siamese network with a unified approach for object classification and bounding box regression. Xu et al. [204] utilized a fully convolutional object tracker with target state estimation and generated classification score for local view instead of predefined anchors. Guo et al. [205] proposed anchor and region proposal free siamese network, which requires fewer hyperparameters. More recently, transformer-based OT has been proposed. Lin et al. [206] proposed the use of a full attention-based siamese network with motion tokens to extract motion context. Cui et al. [207] proposed tracking framework with repetitive mix attention modules for joint features extraction and relation mapping between target and template frame. They also proposed transformer based hierarchical and non-hierarchical trackers with several pretraining strategies. Ye et al. [208] proposed transformer based a single stream stage tracking framework combining feature extraction and target-template relation modeling and utilizing an early candidate elimination approach for efficient inference.

Despite the excellent performance achieved in OT, these trackers still show a comparatively poor result when adverse conditions (e.g. nighttime) are involved. Li et al. [209] proposed a LLE module with a correlation filter for nighttime tracking, but they are restricted to handcrafted features and could not benefit from end-to-end learning. Sasagawa et al. [210] utilized multiple pretrained models for knowledge distillation from the LLE model to the object detection model. Ye et al. [211] proposed retinex based iterative LLE with joint illuminations and noise estimation for OT. They further mitigate the weak collaboration in visual tracking and proposed a transformer [212] with learned curve projection-based image light enhancement for illumination and denoising of low light images. Zhu et al. [213] proposed night image enhancement followed by a tracking approach that utilized darkness clue to produce a visual prompt. Ma et al. [214] proposed a bilevel adaptation for low-light image enhancement, adaptable to unknown scenes.

Chen et al. [215] proposed image and instance level alignment for domain adaptive object detection. Wu et al. [216] used an adversarial learning-based nighttime image relighting module and segmentation module. Ye et al. [217] explored using adversarial learning for

DA in visual OT to minimize the domain discrepancy between day and night features. Their approach has shown promising results in addressing the domain shift problem in OT. Zhang [218] proposed progressive style translation using domain invariant content details and separate domain style details. Yao et al. [219] used a segment-anything model-assisted approach with zero-shot learning-based training sample generation for the target night domain. Wu et al. [220] used an adversarial learning-based single-stage DA approach for semantic segmentation of nighttime images. Lu [221] used a multi-level denoising transformer, which used a self-attentive encoder and decoder with a cross-attention module. Fu et al. [222] proposed a contrastive learning-based domain adaptive network for OT with scale information. Lv et al. [223] used Gabor filter-based preprocessing steps before utilizing the adversarial learning-based domain adaptive training network. Sun et al. [224] utilized filtering of high-frequency noise for nighttime image enhancement with a dynamic template-based OT network. Kennerley et al. [225] proposed a student-teacher-based network that combined high and low-confidence pseudo labels as two-phase consistency with nighttime augmentations. Chen et al. [226] proposed a mean teacher-based network, which utilized assignment-based object identification and LLE for unlabelled target domain inputs. Zheng et al. [227] proposed multi-source DA for satellite videos, which utilized student-based learning with weak re-identification.

2.6 Limitations of existing works

Based on our observation, the drawbacks of the existing literature for each task (SS, CD, IT, UDA) are listed as follows:

2.6.1 Semantic Segmentation

- The field-of-view of the convolution neural network is limited by kernel size.
- Information is lost in the interpolation of features into different lower scales. Preservation of original resolution may help to produce better semantic maps.

- Identification of small objects is a difficult task in highly unbalanced remotely sensed images.

2.6.2 Change Detection

1. The over-fitting is very prominent in the remote sensing CD due to the abundant presence of no-change class pixels.
2. The self-attention based approaches are computationally intensive.
3. Poor and noisy feature representations deteriorate the output of the CD model.

2.6.3 Image Translation

1. Existing methods, including state-of-the-art methods, are unable to capture the finer details and varieties of land-cover information.
2. The convolution kernel has a limited view due to its size, but for precise super-resolution large view is required.
3. For multi-frame super-resolution task, temporal relation needs to be handled.

2.6.4 Unsupervised Domain Adaptation

1. Differences in resolution, sensors, capturing conditions, and inter-class relation between domains are common in remote sensing datasets, which need to be handled effectively.
2. Input level adaptation is difficult due to differences in the styles of domain and target images.
3. The class distribution gap of source and target domain makes the output level adaptation challenging.

4. Learned style translation requires additional computational overhead and may deteriorate results in some unpaired cases.

5. Existing trackers often perform poorly in adverse conditions, such as low-light environments that can affect the visual aspect of the tracked object.

2.7 Summary

This chapter depicted background details of the Remote Sensing Vision Task (RSVT)s (SS, CD, IT, UDA) of the work and presented a detailed literature survey for related works. The limitations of existing methods for each vision applied are highlighted. Limited field-of-view of the kernel, loss in feature interpolation, extensive resource requirement of self-attention, and difficulty of tiny object identification are some challenges faced by existing SS methods. In CD task, existing methods faced class imbalance and noisy feature representation challenges. Existing IT methods suffered the challenges of multi-context view representation and temporal relation mappings. In UDA techniques, challenges of differences in resolution, sensors, capturing conditions, and inter-class relation between domains are common in remote sensing datasets, which need to be handled effectively. With this background and relevant literature, the first contribution of the thesis on SS is presented in the next chapter.

"Remember that wherever your heart is, there you will find your treasure."

~Paulo Coelho, The Alchemist

3

Far pixel relation based deep models for semantic segmentation

This chapter discusses the first contribution to the Semantic Segmentation (SS) task. Image SS involves the classification of each pixel of the input image to a specific class. These specific classes to which input pixels can belong are predefined. SS of remotely sensed data provides accurate delineation of various features (classes) present on the earth's surface, which are required in several analysis and decision-making applications such as natural resource mapping, urban sprawl monitoring, deforestation tracking, disaster surveillance, etc. In this chapter, we proposed two Deep Learning (DL) models named Aggregated Context Network (ACNet) and HybridNet for enhanced SS. The significant contributions

of this chapter are summarized as follows:

- The proposed ACNet model utilized attention-based skip connection, dilated convolution extracted multi-context, and graph convolution-based far pixel relation mapping with hard negative mining for enhanced SS.
- The proposed HybridNet model further improved the results for SS with combined global pixel interaction, dense dilated aggregated field of view, though voting and self-attention.

3.1 ACNet

SS of Very High Resolution (VHR) imagery is an important task where context information plays a crucial role. Adequate feature delineation is difficult due to high-class imbalance in remotely sensed data. Maggiori et al. [149] analyzed various DL architecture for segmentation tasks and proposed multilayer perceptron network. Encoder-decoder based networks [5,228] have shown significant performance in SS tasks due to their fully convolutional nature, which allows preservation of spatial information. Li et al. presented Scattnet [229]. The authors have proposed an end-to-end deep SS network, integrating lightweight spatial and attention modules to refine the features adaptively. It is observed that the existing SOTA methods for SS of VHR aerial imagery class imbalance is a significant problem in SS. In aerial data, this problem becomes more prominent. It is challenging to generate the correct segmentation mask for the least represented classes.

In ACNet, we proposed a variant of encoder-decoder-based architecture with residual attentive skip connections and multi-context blocks to capture multi-scale and multi-context features. The major contributions in ACNet are as follows:

1. We proposed a dilated multi-context block in every encoder layer to capture a larger field of view.
2. An attentive skip connection is incorporated to capture multi-scale skip information more effectively between the encoding and decoding pipeline.

3. We used graph convolution based long-range context extraction for better semantic map generation.

4. Online hard negative mining based training strategy is utilized to handle class imbalance.

3.1.1 Proposed Network Architecture

The overall architecture of ACNet is depicted in Fig 3.1. The channel-wise attention and spatial attention are incorporated through the skip connections. Global Context Module (GCM) is utilized to capture long-range dependency between pixels. Individual modules are described below:

3.1.1.1 Encoder Block

We proposed a multiple encoder block (EB) to capture multi-scale features. Encoder block consists of a multi-context block implemented using dilated convolution and dense block as shown in Fig. 3.1. A multi-context block enables a larger field of view as input is processed by three parallel convolutions with a dilation rate [1,2,4]. This can be represented as

$$O_{mc} = conv_{d_1}(x) + conv_{d_2}(x) + conv_{d_4}(x) \tag{3.1}$$

Here x is input to encoder block, O_{mc} is output of it and $Conv$ is convoltion operation with dilation rate d_i. For efficient feature representation, these features are combined by addition operation and passed to the dense block of four layers. Last layer of EB used the pooling layer to down-sample the features space.

3.1.1.2 Decoder Block

The decoder block (DB) upsampled the features of the corresponding EB block using a deconvolution layer and combined it to residual attentive skip connection (RAC) using addition operations. These features are passed to the sequence of convolution, batch normalization,

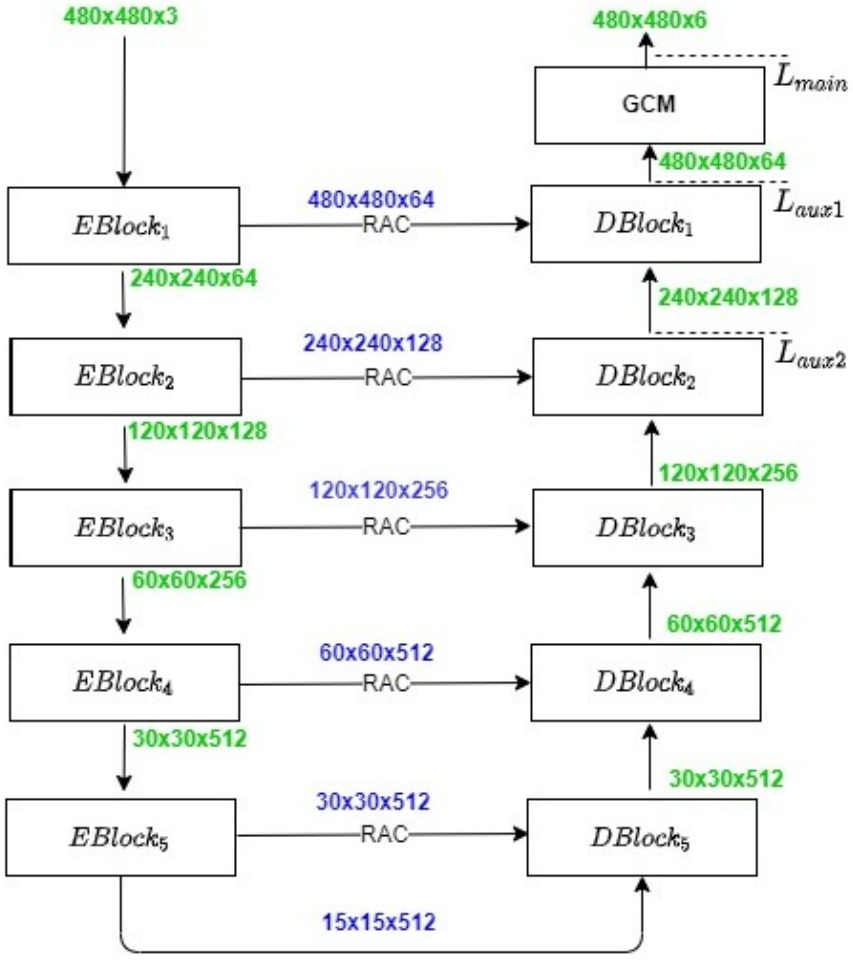


Figure 3.1: *Our Encoder-Decoder Architecture; Here E block denotes encoder block, D Block denotes decoder block, GCM is Global Context Module, L_{main} is the main loss, L_{aux1} and L_{aux2} are auxiliary losses, the dimension of input and output of a block has been shown in lime color, the dimension of skip connection has been shown in aqua color*

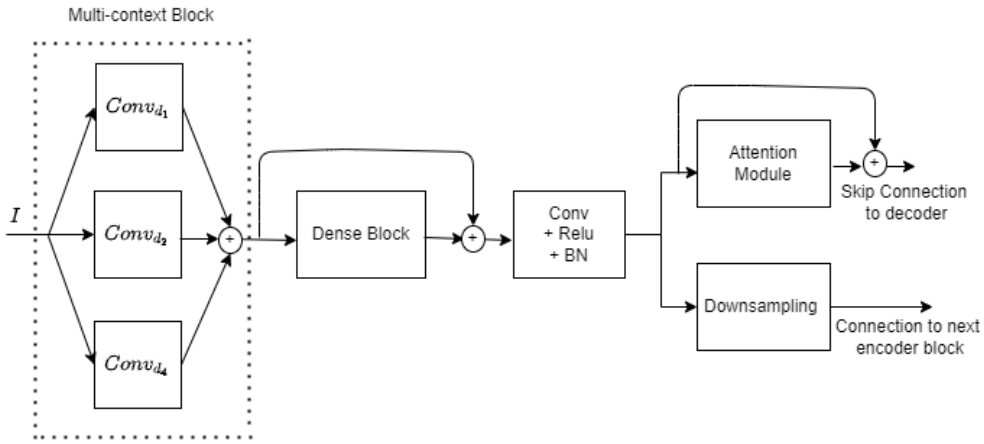


Figure 3.2: *Proposed Encoder Block Architecture; Here* I *is input of the network, $conv_{d_i}$ is convolution with dilation rate d_i, Conv is convolution operation,BN is batch nomralization layer.*

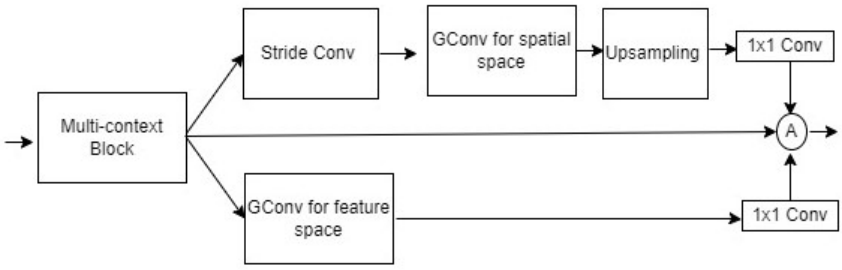


Figure 3.3: *Global context Module, here GConv is graph convolution, conv is convolution layer, and, A denotes sum operation.*

	ImSurface F_1	Building F_1	Low Vegetation F_1	Tree F_1	Car F_1	OA
FCN-8s [4]	80.10	85.17	72.77	68.05	81.67	77.82
UNet [5]	79	84.82	78.87	74.91	87.47	79.84
SegNet [228]	81.07	86.41	77.98	73.85	85.72	80.27
RefineNet [230]	87.58	88.50	81.91	79.07	87.90	84.38
DeepLab [231]	89.31	92.81	83.37	78.40	88.24	86.76
G-FRNet [232]	88.20	92.69	82.84	79.00	86.28	86.84
SCAttNet [229]	90.04	90.05	84.05	79.75	89.06	87.97
ACNet	91.39	95.46	84.65	79.82	97.58	89.16
ACNet on er	**93.14**	**96.05**	**86.77**	**84.46**	**95.29**	**91.03**

Table 3.1: *Potsdam dataset evaluation results in comparison on the test set. Here bold represents the best result, and er is the results evaluated on eroded labels*

	ImSurface F_1	Building F_1	Low Vegetation F_1	Tree F_1	Car F_1	OA
UNet [5]	88.57	90.06	78.81	80.08	68.74	84.88
SegNet [228]	86.19	88.32	77.17	79.09	54.24	82.92
RefineNet [230]	87.27	86.90	78.34	79.63	75.89	83.36
DeepLab [231]	88.54	89.95	80.15	80.05	68.91	85.15
G-FRNet [232]	88.94	90.47	80.13	80.46	72.60	85.52
SCAttNet [229]	89.13	90.30	80.04	80.31	70.50	85.47
ACNet	89.61	90.77	81.92	83.83	69.63	86.72
ACNet on er	**93.03**	**93.46**	**85.44**	**86.15**	**86.28**	**90.07**

Table 3.2: *Vaihingen dataset evaluation results in comparison on the test set. Here, bold represents the best result, and er is our results evaluated on eroded labels*

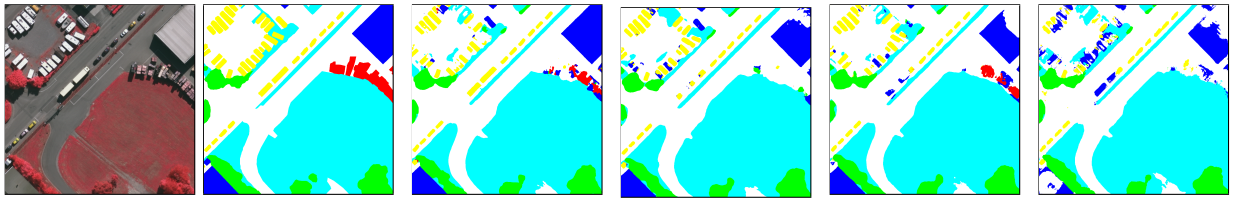
and Relu layers. DB performed reconstruction of encoded features and feature refinement in higher resolutions.

3.1.1.3 Residual Attentive skip connections

As discussed in the previous subsection, the EB is connected with the DB through residual attentive skip connections (RAC). It combines the low and middle-level features extracted through EB with the middle and high-level features of DB. RAC incorporates an attention module to focus on the most relevant features. These RACs use convolutional block attention (CBAM) [233] for the attention map. CBAM contains the channel attention module and spatial attention module.

The dimension of input (F), spatial attention, map and channel attention map are denoted by $F\ \epsilon\mathbb{R}^{\mathrm{Ch\ x\ H\ x\ W}}$, $A_{\mathrm{s}}\ \epsilon\ \mathbb{R}^{1\ \mathrm{x\ H\ x\ W}}$ and $\mathrm{A_c}\ \epsilon\ \mathbb{R}^{\mathrm{Ch\ x\ 1\ x\ 1}}$, respectively. Here H, W, Ch are height, width, and the number of features in the input. The spatial attention map (A_{s}) and channel attention map (A_{c}) are combined with input (F) to generate final attentive map F_{ii} as per following equation:

$$\begin{aligned} F_{\mathrm{i}} &= A_{\mathrm{c}}(F) \otimes F \\ F_{\mathrm{ii}} &= A_{\mathrm{s}}(F_{\mathrm{i}}) \otimes F_{\mathrm{i}} \end{aligned} \tag{3.2}$$



(a) *Input* **(b)** *GT* **(c)** *Final* **(d)** *DB+RAC+CB* **(e)** *DB + RAC* **(f)** *DB*

Figure 3.4: 1024×1024 *pixels patches output comparison, Here DB is our base network with dense connection, RAC is residual attentive skip connection, CB is context block. More results are included in the supplementary material.*

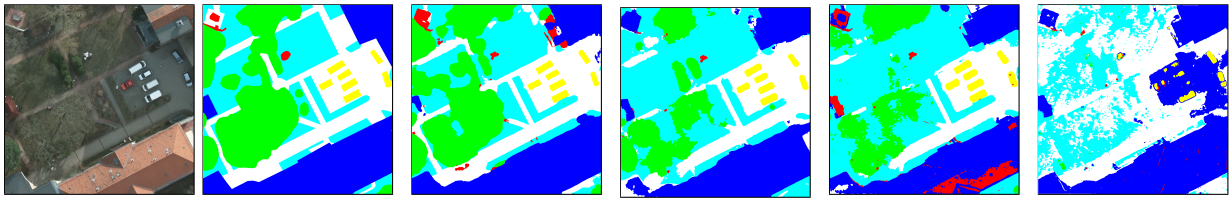
Here $\otimes$ is the pointwise multiplication operator. Channel attention (A_c) is calculated using channel-wise average and max pooling operation with multilayer perceptron. Spatial attention (A_s) is calculated using spatial max and average pooling with a convolution filter of the kernel of 7×7. The channel and spatial attention modules emphasize regions of interest concerning the input image's object and object locations.

3.1.1.4 Global Context Module

In the proposed architecture, the series of DB is followed by GCM as the last module. GCM is used to capture long-range dependency between pixels at the image's full scale. It consists of a multi-context block to capture multi-view features, and graph convolution networks as used in [234]. It is noted that the GCM block's output is the output of the overall proposed model architecture. Graph convolution can be defined as

$$G_c = \sigma(MIW) \tag{3.3}$$

Here G_c is graph convolution, M is adjacency matrix contains graph connectivity information, W is weigh matrix, σ is activation function, and I is input feature. As shown in Fig. 3.3, two graph convolution operation one in spatial space and another in feature space is utilized in this module.



(a) *Input* (b) *GT* (c) *Final* (d) *DB+RAC+CB* (e) *DB + RAC* (f) *DB*

Figure 3.5: 1024×1024 *pixels patches output comparison, Here DB is our base network with dense connection, RAC is residual attentive skip connection, CB is context block. More results are included in the supplementary material.*

3.1.2 Experiments

3.1.2.1 Dataset

In this work, the proposed model's performance is evaluated on *ISPRS Vaihingen* and *ISPRS Potsdam* benchmark datasets. The dataset is divided into train and test set for evaluation, as mentioned in [229]. We used both normal and eroded boundaries ground truth labels for performance comparison.

3.1.2.2 Training

We have generated 6000 random patches of 480x480 pixels for each data set. The model is trained for total epochs of 150 with a batch size of 4. The stochastic gradient descent (SGD) optimized with initial learning rate 0.001, momentum value 0.9, and weight decay value 0.0005 is used for training. We have reduced the learning rate by 0.5 whenever the model is unable to reduce the mean loss for 15 epochs. *Categorical cross entropy* loss is used as the loss function, which is composed of softmax activation and cross-entropy loss. The Loss function is defined as

$$Loss = \lambda_1 \times L_{main} + \lambda_2 \times L_{aux1} + \lambda_3 \times L_{aux2} \tag{3.4}$$

Here, L_{main} is main loss function, and L_{aux1}, L_{aux2} are two auxiliary loss functions as used in [235]. After experiments, we found that $OHEM$ ratio of 0.7 and λ_1, λ_2, λ_3 values of [1.0, 0.4, 0.2] gives optimal results.

3.1.3 Results

We have used *F1 score* (F_1), *overall accuracy (OA)*, and mean *intersection over union (mIoU)* as the performance measures of our model. The formula for calculation of F_1 is

$$FS = 2 \times \frac{P_r \times R_c}{P_r + R_c} \tag{3.5}$$

Here Precision (P_r) and Recall (R_c) can be calculated based on true positive(t_p), false positive (f_p), true negative (t_n), and false negative(f_n). *OA* is calculated as

$$OA = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \tag{3.6}$$

Visual comparison of the output of various components for test sets are shown in Fig. 3.4 and Fig. 3.5. Quantitative comparison with other published work is presented in Table 3.1 and Table 3.2. Our model outperformed other works in classwise *FS* and *OA*.

Our Method	Parameters	OA
Base	34 M	79.86
Base + DE	32.81 M	81.95
Base + DE + RAC	32.89 M	82.88
Base + DE + RAC + CB	41.92 M	85.43
Base + DE + RAC + CB + GCM	42.13 M	**86.72**

Table 3.3: *Ablation Study on* Vaihingen *test dataset, Here Base is a base network, DE is a dense encoder, ARS is attentive residual skip connection, CM is Context-module, GCM is global context module.*

3.1.4 Ablation Study

In Table 3.3, a detailed ablation study is presented. We run all our ablation experiments on the *Vehingen* dataset, and matrices are calculated on the test set. We started with an initial model named Base, which is based on encoder-decoder architecture and achieved *OA* of 79.86%. We enhanced the encoder units with the replacement of the convolution sequence with the dense blocks, which resulted in an improvement of 2.09% of *OA*. This also reduces the number of trainable parameters. We enhanced the skip connection with the inclusion of attention units and achieved *OA* of 82.88%. Contextual information plays an important role in precise semantic map delineation. To incorporate this, we added a context block in the network that resulted in *OA* of 85.43%. Finally, we added GCM to handle long-range pixel interactions and achieved a final *OA* of 86.72%.

3.2 HybridNet

Among DNN based SS models, encoder-decoder-based architectures [4–11] are popular due to their simplicity and extendability. The preservation of the original resolution of input [12–15], is also frequently reported. The major disadvantage of these approaches is that they are primarily computational intensive and require higher graphics memory. It is also observed that the field of view plays a vital role in SS. Thus the large kernel size is a natural choice to increase the effective field of view of the filter [16, 236]. Alternatively, few approaches use atrous convolution [7, 11, 17–19], which utilizes a hole in between the kernels to result in a large kernel effect. The large kernel achieves a large field of view without increasing resource requirements. Attention is a mechanism to highlight the important features and utilized in recent work [18, 20–22]. Dense connection based neural networks [19, 23] provide parameter sharing and produce better results forSS. A conditional random field is used by some approaches [24, 25] as a post processing for enhancement of the segmentation results. Super pixels with Convolutional Neural Network (CNN) based approach [26] is also proposed for effective SS. Some recent approaches use the generalized hough transform to find distant pixel interaction [27–31]. The current works for remote sensing image segmentation are mainly based on an encoder-decoder architecture that has a limitation of intrinsic information loss due to continuous downsampling operation. DNN-based work also faces challenges in extracting the inputs' local and global context due to kernel size and computational limitations. The remote sensing datasets also have a class imbalance.

In HybridMet, We extend a HR network with dense connection integration, lightweight self-attention module, and generalized hough transform based deep voting module for SS. The major contributions to HybridNet are as follows:

1. We proposed a novel multi-scale architecture HybridNet with Combined Global Pixel Interaction (CGPI) approach for effective SS.

2. Dilated convolution based densely connected aggregated field-of-view module (AFM) is proposed to capture multi-context features from the input.

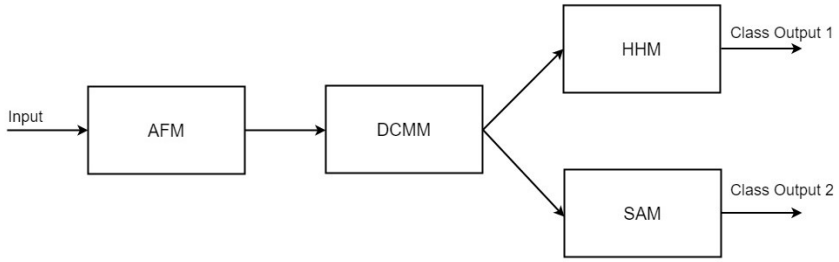


Figure 3.6: *Proposed architecture, Here AFM is aggregated field-of-view module, DCMM is densely connected multi-resolution module, HHM is hough voting module, and SAM is self attention module.*

3. We introduced a densely connected multi-resolution module (DCMM) as a backbone network to achieve better feature representation.

4. We incorporated the heatmaps generated by log-polar based hough voting in the hough heatmap module (HHM) of the proposed model for the SS task.

5. A lightweight spatial self-attention module is utilized for adequate delineation of objects.

3.2.1 Proposed Network Architecture

The proposed HybridNet architecture has four sub-module, namely Aggregated Field-of-view Module (AFM), Densely connected Multi-resolution Module (DCMM), Hough Heatmap module (HHM), and Self Attention module (SAM) as described in Fig. 3.6. Input is passed to AFM for multi-context feature extraction, which is forwarded to DCMM for detailed feature generation at multiple scales. These features are parallelly passed to SAM and HMM for far pixels dependency extraction. The network structure details of input and output shapes of various modules are shown in Table 3.5. The CNN parameters of the initial channel are selected based on a comparative study between 8, 16, 32, and 36 channels. The loss weight parameter value of 0.3 is found out after the grid search. The internal number of channels needs to be the same as we have used multi-resolution dense connections.

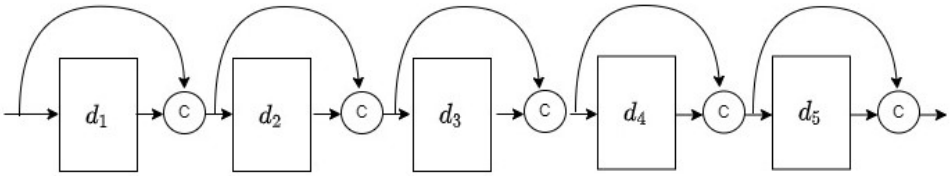


Figure 3.7: *Aggregated Field-of-view module module. Here, c is a concatenation operation, and each block is convolution with a dilation rate of d_i.*

3.2.1.1 Aggregated Field-of-view Module

We proposed an aggregated filed-of-view module (AFM), as shown in Fig 3.7. AFM is based on multiple dilated convolutions and dense connections. The multi-context features are extracted from the input image to provide enhanced field-of-view using this module. The dilation operation can be expressed as

$$(F *_{\mathrm{d}} K)(p) = \sum_{s+dt=p} F(s)K(t) \tag{3.7}$$

where $*_{\mathrm{d}}$ is dilated convolution operation with dilation rate d, $(F *_{\mathrm{d}} K)(p)$ is output, $F(s)$ is input, and $K(t)$ is kernel used. In this sub-module, we used a sequence of increasing dilation rates of [1, 2, 4, 8, 16]. Features of each dilated convolution are concatenated with the input of it. The final output of i^{th} unit of AFM can be represented as

$$D_i = conc(d_i(D_{i-1}), D_{i-1}) \tag{3.8}$$

Here, D_i is output of i^{th} unit of AFM, $conc$ is concatenated operation, d_i is dilated convolution of i^{th} unit. The base condition for the above equation with input to the AFM module is x, which can be represented as

$$D_1 = conc(d_1(x), x); D_0 = x \tag{3.9}$$

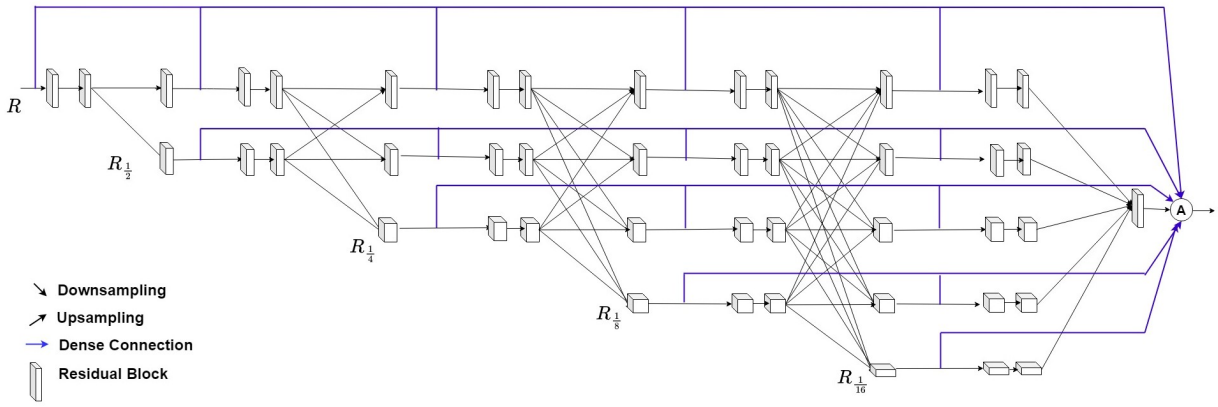


Figure 3.8: *Densely connected multi-resolution module. Here each connection is combined by addition operation, dark blue connection represents proposed dense connection between stages. Black down arrow represent downsampling operation, Black up arrow is upsampling operation. Here R represents the input's resolution, which decreased to one-half of the value in the following resolution streams.*

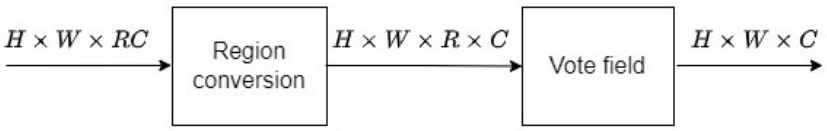


Figure 3.9: *Hough heatmap module. Here, c is concatenation operation and each block is convolution with dilation rate of d_i.*

3.2.1.2 Densely Connected Multi-resolution Module

The multi-context features obtained in the AFM are passed through DCMM, which helps parameter sharing and quality mixing among different contextual feature maps. The concept of DCMM is taken from HRNet [14] which introduced HR connections. As depicted in Fig. 3.8, in the proposed model, we have used densely connected five stages of sub-networks. This architecture essentially consists of five parallel convolution sequences with the resolution of 1, 1/2, 1/4, 1/8, 1/16. We added dense connections between stages of the HR module. The output of the DCMM is passed through the HHM and SAM in a parallel fashion (Please refer to Fig. 3.6). Both the modules produce final class probabilities.

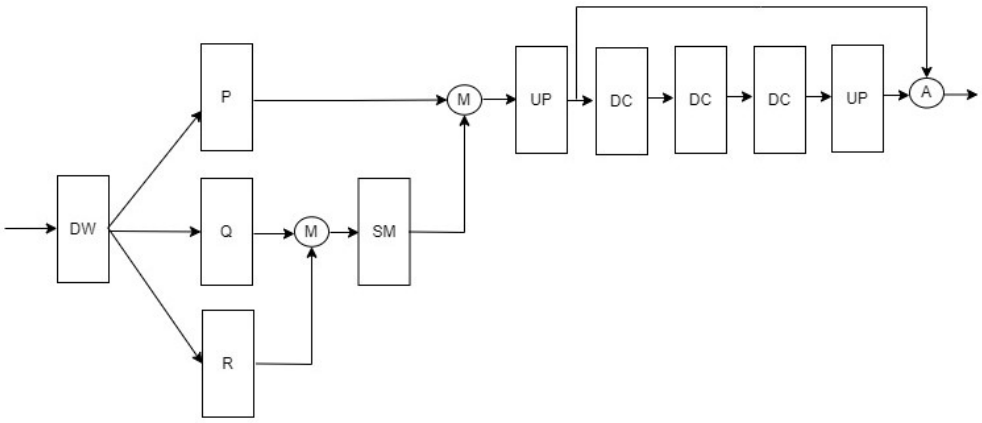


Figure 3.10: *Self attention module. Here, DW is downsampling operation, M is multiplication operation, SM is softmax operation, A is addition operation, UP is upsampling operation, and DC is downsampling through convolution. Three convolution operation is represented by P, Q, R.*

3.2.1.3 Hough Heatmap Module

For long-range dependency, the HHM is incorporated as shown in Fig 3.9, which is based on an existing work [27] where a polar coordinate system is used to create different regions. It [27] proposed Hough voting for the object detection task. We extended it for the SS task. The major difference from the existing works is that we used HHM as an additional pixel classifier utilized during the training phase to enhance feature representation. We did not use hough voting features at inference time for the final segmentation map generation. From each region, the vote field has been calculated. The output of DCMM of dimension $H \times W \times RC$ is passed to HHM. It first transformed the input to $H \times W \times C \times R$ dimension using region conversion submodule. These C tensors of size $H \times W \times R$ are utilized in HHM for voting using the vote field. The HHM produces a C output vote map of size $H \times W$. For each location (i; j) in this 2D map, visual evidence map is generated using deconvolution operation. The polar coordinate system is designed by dividing the input into five concentric circles with [2, 8, 16, 32, 64] radii. These circles are divided into 17 regions each of 90°angles. We used a field size of 65.

3.2.1.4 Self Attention Module

Context modeling is useful in pixel-based segmentation. As pixels are highly co-related in the local neighborhood, local context is useful for understanding the pixel similarity in the neighboring zones. In addition, the Global context extracts long-range dependencies of pixels and provides an overall idea about the input. We used a non-local operator to calculate the affinity matrix of the input. A SAM inspired by [20, 21] is used for the global context of features as shown in Fig. 3.10. We generated self-attention maps at $\frac{1}{4}^{th}$ of the original resolutions of the input to reduce the computational requirement. It consists of self-attention followed by local aggregation. Self-attention is implemented using three convolutions (P, Q, R) and $softmax$ and multiplication operations that can be represented as

$$A = softmax(Q \times R) \times P \tag{3.10}$$

Global context aggregated features extracted by self-attention are further refined using the sequence of learned downsampling using convolution followed by non-learned bi-linear upsampling operation. It can be represented as

$$A_f = A + \sigma(Up(ConvD(ConvD(ConvD(A))))) \times A \tag{3.11}$$

Here, A_f, A, σ, $ConvD$, and Up represent the final feature from the SAM, Output of self-attention, sigmoid activation function, downsampling operation with convolution and upsampling operation.

3.2.2 Experiments

3.2.2.1 Dataset

In our experimental setting, we have used *ISPRS Potsdam* and *ISPRS Vaihingen* datasets. These are benchmark datasets and are heavily used for performance comparison of aerial SS tasks. The *ISPRS Potsdam* dataset images are of size 6000 × 6000 pixels. The *ISPRS*

Evaluation	Dataset	Train set ids	Test set ids
Local	*Vaihingen*	1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37	11, 15, 28, 30, 34
Benchmark	*Vaihingen*	1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, 37	2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38
Benchmark	*Potsdam*	2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_10, 7_11, 7_12	2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13

Table 3.4: *The dataset splits used in our experiments for local and benchmark evaluations.*

Vaihingen dataset images are of varying sizes in the range of 1900 -3000 pixels in rows and columns. To showcase the efficacy of the proposed model, we compared our result with other SOTA schemes as reported in local evaluation [9, 149, 237, 238] and benchmark evaluation [11, 19, 146]. The details about image ids used in the different split settings are given in Table 3.4. In our experiments, we use three band architecture that utilized Red-Green-Blue bands for *ISPRS Potsdam* and Infrared-Red-Green for *ISPRS Vaihingen* datasets. These datasets contain semantic annotations for six classes as mentioned in Tabel 3.6. Fair evaluations are guaranteed using the same train and test dataset settings as other SOTA methods in local evaluation [9, 149, 237, 238] and benchmark evaluation [11, 19, 146].

3.2.2.2 Training

We proposed a model with a sufficiently large number of parameters and trained it for many epochs to avoid underfitting. To handle overfitting, we use data augmentation, early

Module	Input Shape	Output Shape
AFM	$8 \times 3 \times 256 \times 256$	$8 \times 32 \times 256 \times 256$
DCMM	$8 \times 32 \times 256 \times 256$	$8 \times 256 \times 256 \times 256$
HHM	$8 \times 32 \times 256 \times 256$	$8 \times 6 \times 256 \times 256$
SAM	$8 \times 32 \times 256 \times 256$	$8 \times 6 \times 256 \times 256$

Table 3.5: *The network structure of the HybridNet. Here, shapes are presented in BCHW format where B is batch size, C is channels, H height, and W is width of input.*

stopping in during training. We started with the local evaluation of the model for *ISPRS Vaihingen* dataset and used the data split as mentioned by [9, 149, 237, 238]. We generated 6000 random patches of size 256×256 pixels from the train set and trained the model. We used both datasets individually in benchmark evaluation and followed the data split as used in [11, 19, 146]. We generated 6000 random patches of size 256×256 pixels for both datasets to train our model from the respective train sets. We trained the model for a total of 300 epochs for the respective dataset in each evaluation. The proposed model is implemented using PyTorch [239] library. We used values of 8, 0.0001, and 0.00005 for batch size, initial learning rate, and weight decay in all experiments. The batch and patch sizes are selected based on the available graphics card memory resource. The initial learning rate and batch decay values are chosen based on grid search. We used a batch normalization layer in our model.

$$Loss = L_{main1} + L_{main2} + 0.3 \times L_{aux} \tag{3.12}$$

Eqn 3.12 represents the loss function for the proposed model. Here, L_{main1} is loss calculated on attention module output, L_{main2} is loss calculated on hough module output, and L_{aux} is auxiliary loss. We used Categorical *Cross-entropy* loss with online hard negative mining in all loss functions. It is the combination of softmax activation and cross entropy loss. We used the softmax activation function in the final layer of the network. Softmax function f can be expressed as

$$f(s)_i = \frac{e^{s_i}}{\sum_{j}^{C} e^{s_j}} \tag{3.13}$$

It represents the probability of input s belonging to class i. Here, C is the total number of

classes. *Cross entropy*(CE) over softmax can be expressed as

$$CE = -\sum_{i}^{C} t_i \log(f(s)_i) \tag{3.14}$$

Here, CE is cross-entropy, t is the target vector, and its value will be 1 when ground truth belongs to class i. We performed the Almost Stochastic Order test [240,241], and the Almost Stochastic Dominance test [242] for statistical analysis of the proposed model compared to other SOTA models.

3.2.2.3 Evaluation Parameters

We used standard parameters of *F1 score* (F_1), *overall accuracy (OA)*, and *mean intersection over union (mIoU)* for the evaluation of our model.

3.2.3 Results

3.2.3.1 Local Evaluation

In Table 3.7, we presented the result of our model for local evaluation of the *ISPRS Vaihingen* dataset. We achieved OA and F_1 of 89.91% and 89.75 %. We got the improvement of 0.81 % and 0.2 % in F_1 and OA values from the recently published SOTA [237]. In local evaluation, we obtained the best F_1 values for low vegetation, tree, and car classes and the second best F_1 values for impervious surface and building classes. The best F_1 for impervious surface and building classes are produced by HCANet [237] with ResNet-101 encoder. We achieved significant improvement in F_1 and OA values. HCANet [237] used a large patch size of 512×512 with multi GPU setting. Our model can surpass OA, F_1, and most of class F_1 with a small patch size of 256×256 with a single GPU.

3.2.3.2 Benchmark Evaluation

In Table 3.8 and Table 3.9, we presented benchmark evaluation results. We achieved the best results of $OA, mIoU$ for *Vaihingen and Potsdam* dataset with values of F_1 and OA of 90.80 % and 91.10 %. It improved the F_1 by 0.7% and OA by 0.2% from the previous

Class Name	Color Code
Building	
Impervious surfaces	
Low vegetation	
Tree	
Car	
Clutter/background	

Table 3.6: *Color code used in groundtruth images of Vaihingen and Potsdam SS datasets.*

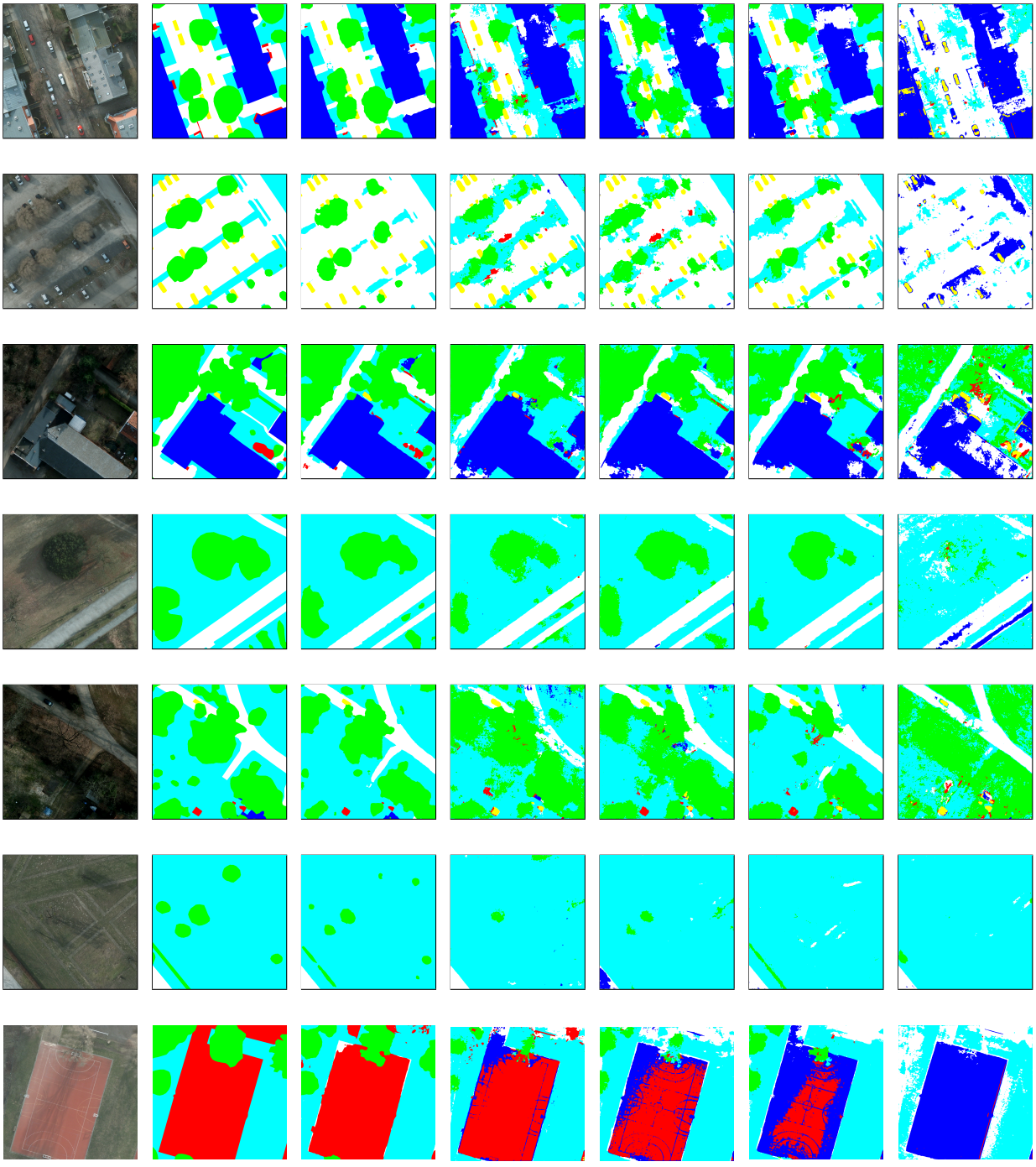
Models	ImSurface	Building	Low Veg.	Tree	Car	Avg. F_1	OA	mIoU
Rot-EQ-Net [9]	89.50	94.80	77.50	86.50	72.60	84.18	87.50	-
FCN-FR [149]	91.69	95.24	79.44	88.12	78.42	86.58	88.92	-
S-RA-FCN [238]	91.47	94.97	80.63	88.57	87.05	88.54	89.23	-
HCANet Res-34 [237]	91.61	95.22	80.24	88.91	86.56	88.51	89.39	
HCANet Res-101 [237]	92.20	95.55	80.66	88.92	87.36	88.94	89.71	
Proposed	92.09	95.45	81.68	89.17	90.39	89.75	89.91	68.10

Table 3.7: *Vaihingen local evaluation result comparison. Here, red represents the best result, and blue represents the second-best result.*

work [144]. We got the best F_1 for three classes of low vegetation, tree, and car and the second best F_1 for impervious surface and building classes. For *Potsdam* dataset we achieved F_1 and OA values of 92.94 % and 91.53 %. It is the improvement of 0.04% and 0.03 % in F_1 and OA over previous SOTA results [11]. We got the best F_1 values for low vegetation and car classes and the second best F_1 results for impervious surface and tree classes. We obtained $mIoU$ values of 84.30 % and 86.88 % respectively for *Vaihingen* and *Potsdam*.

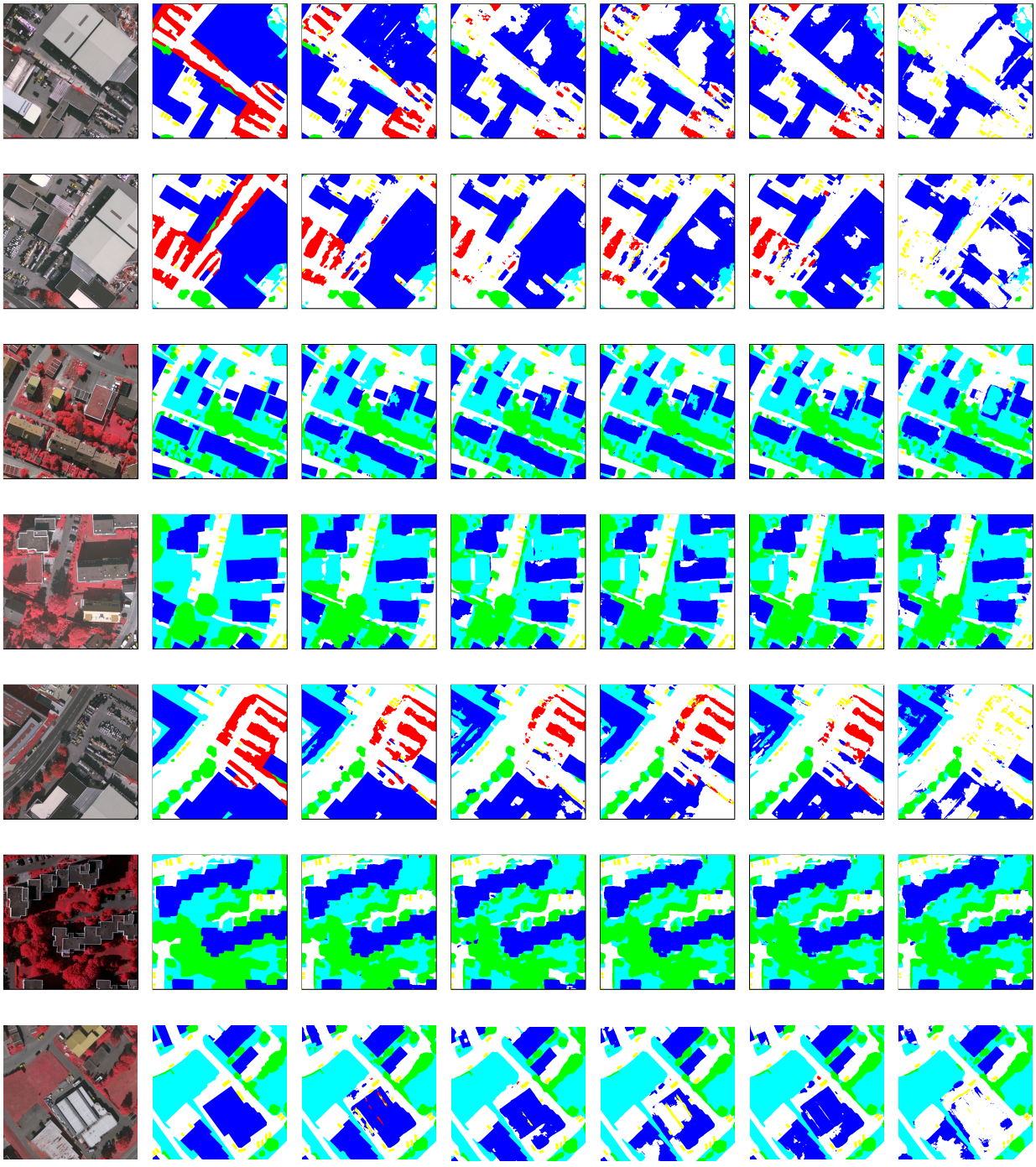
3.2.4 Ablation Study

In Table 3.11, we presented a detailed ablation analysis of the network. We showed the result of *Vaihingen* benchmark test dataset in all experiments of ablation studies. We started with the baseline model represented as Base, which is modified HRNet [14] with 32 initial channels. It achieved the OA of 89.79% and $mIou$ of 80.83. We added a hough module with Base, which improved 0.31% in OA and 1.46% in $mIoU$. The combination of the base model and attention module produced an improvement of 0.7% in OA and 1.48%



(a) I (b) GT (c) F (d) $_{B+A+S}$ (e) $_{B+A+H}$ (f) $B+A$ (g) B

Figure 3.11: 1024×1024 *pixels patches output comparison of Potsdam dataset. Here I is input, GT is ground truth, F is the final output, B is our base network with dense connection, A is aggregated field-of-view module, H is the hough heatmap module, and S is a self-attention module. Here, the first column is the input image which is followed by ground truth, the output of the final network, the output of Base + AFM + SAM network, the output of Base + AFM+ HHM network, the output of Base + AFM, and the output of Base network.*



(a) I (b) GT (c) F (d) $_{B+A+S}$ (e) $_{B+A+H}$ (f) $B+A$ (g) B

Figure 3.12: 1024×1024 *pixels patches output comparison of Vaihingen dataset. Here I is input, GT is ground truth, F is the final output, B is our base network with dense connection, A is aggregated field-of-view module, H is the hough heatmap module, and S is a self-attention module. Here, the first column is the input image which is followed by ground truth, the output of the final network, the output of Base + AFM + SAM network, the output of Base + AFM+ HHM network, the output of Base + AFM, and the output of Base network.*

Models	ImSurface	Building	Low Veg.	Tree	Car	Avg. F_1	OA	mIoU
INR [149]	91.69	95.24	79.44	88.12	78.42	86.58	88.92	-
CRDNet [243]	92.70	95.40	83.40	89.60	88.70	90.50	90.00	-
V-FuseNet [148]	92.00	94.40	84.50	89.90	86.30	89.42	90.00	-
DLR_9 [147]	92.40	95.20	83.90	89.90	81.20	88.52	90.30	-
TreeUNet [146]	92.50	94.90	83.60	89.60	85.90	89.30	90.40	-
DDCM [19]	92.70	95.30	83.30	89.40	88.30	89.80	90.40	81.70
MANet [244]	93.02	95.47	84.64	89.98	88.95	90.41	90.96	82.70
A2FPN [245]	93.00	95.70	84.70	90.00	86.90	90.10	91.00	82.20
CASIA2 [144]	93.20	96.00	84.70	89.90	86.70	90.10	91.10	-
Proposed	93.26	96.02	85.13	89.95	89.62	90.80	91.30	84.30

Table 3.8: *Vaihingen benchmark evaluation result comparison. Here, red represents the best result, and blue represents the second-best result.*

Models	ImSurface	Building	Low Veg.	Tree	Car	Avg. F_1	OA	mIoU
V-FuseNet [246]	92.70	96.30	87.30	88.50	95.40	92.04	90.60	-
TreeUNet [146]	93.10	97.30	86.60	87.10	95.80	91.98	90.70	-
CRDNet [243]	92.90	96.70	87.40	88.60	94.80	92.10	90.70	-
DDCM [19]	92.90	96.90	87.70	89.40	94.90	92.30	90.80	86.00
CASIA3 [144]	93.40	96.80	87.60	88.30	96.10	92.44	91.00	-
RS+Edge+DNL [247]	94.00	97.10	87.90	87.60	96.30	92.30	91.10	-
A2FPN [245]	93.60	96.90	87.50	88.40	95.70	92.40	91.10	86.10
BKHN3 [145]	93.30	97.20	88.00	88.50	96.00	92.60	91.10	-
MANet [244]	93.40	96.96	88.32	89.36	96.48	92.90	91.32	-
ResUNet-a d7v2 [11]	93.50	97.20	88.20	89.20	96.40	92.90	91.50	-
Proposed	93.69	97.10	88.48	88.91	96.43	92.94	91.53	86.88

Table 3.9: *Potsdam benchmark evaluation result comparison. Here, the red color represents the best result, and the blue color represents the second-best result.*

Module	OA
Boundary Loss [248]	90.59
Super Pixel loss [249]	90.67
CE	91.01
RMI [250]	91.07
CE + OHEM	91.11
CE + OHEM + Aux	91.14

Table 3.10: *Different loss function's performance comparison*

Our Method	PM	GF	OA	mIoU
Base	27.21M	9.54G	89.76	80.83
Base + HHM	27.22M	9.61G	90.07	82.29
Base + SAM	27.25M	9.68G	90.46	82.31
Base + SAM + HHM	27.26M	9.75G	90.56	82.79
HBase + SAM	38.68M	86.65G	90.65	83.05
HBase + SAM + AFM	38.77M	92.55G	90.78	83.36
HBase + SAM + HHM + AFM	38.77M	92.60G	90.91	83.66
HBase + SAM + HHM + AFM + Dense	38.78M	93.60G	91.14	83.97
HBase + SAM + HHM + AFM + Dense + DA	38.776M	93.60G	91.30	84.30

Table 3.11: *Ablation Study on* Vaihingen *test dataset, Here Base is a base network, SAM is self-attention module, AFM is aggregated field-of-view module, HHM is hough heatmap module, HBase is modified base network, Dense is dense connectivity in HBase, and DA is test time augmentation. PM is the number of parameters, and GF is the gigaflops required by different model configurations.*

in $mIoU$. We combined all three modules, resulting in OA of 90.56% and $mIoU$ of 82.79%. We introduced skip connections in a base module, which is represented by HBase. Combined with the attention module, it produced the OA of 90.65% and $mIoU$ of 83.05%. Integration of AFM before HBase-32 + att improves the performance further to 90.78% for OA and 83.36% for $mIoU$. Combining this with the Hough module resulted in OA of 90.91% and $mIoU$ of 83.66%. Finally, we added dense connectivity in place of skip connections within HBase-32, which further improved the overall accuracy to 91.14%. We performed test time augmentations in the final evaluation and achieved a final OA of 91.30% and $mIoU$ of 84.30%.

Visual comparison of the output of the ablation study is presented in Fig. 3.12 and Fig. 3.11. The base model with dense connection identified the building almost correctly but missed most of the background objects for both datasets. The background objects are classified as cars, buildings, and other pixels. The use of the attention module improved the background classifications but enhanced the impervious detections. Using HMM improved the object boundaries and further reduced the confusion between objects. With self-attention, the prediction map became less confusing. In the final model, the minor object of the car can be identified correctly, and other objects' boundaries are also detected, similar to ground truth. It proved our hypothesis that using two long-range pixel dependency approaches can improve the quality of generated segmentation maps.

We experimented with different loss functions for SS tasks using *Vaihingen* test dataset. A detailed study is presented in Table 3.10. We started with boundary loss [248], and the model achieved OA of 90.59%. Superpixel loss [62] improved the OA by 0.08%. With the cross-entropy loss, we achieved OA of 91.01%, which further improved to 91.07% using RMI loss [250]. Cross entropy with online hard negative mining resulted in OA of 91.11%, which further improved to 91.14% using additional auxiliary loss.

The model's computational complexity is also shown in Table 3.11. The Base required 27.21 million parameters and 9.54 gigaflops. Adding hough modules slightly increases 0.01 million and 0.07 gigaflops in parameters and floating-point operation requirements. Base and attention modules required 27.25 million parameters and 9.68 gigaflops. All three modules required 27.26 million parameters and 9.75 gigaflops. The modified Base named HBase with attention module required 38.68 million parameters and 86.65 gigaflops. The addition of AFM increased the parameters and flops requirement by 0.09 million and 5.9 gigaflops. After the inclusion of HHM, the resource requirement came to 38.77 million parameters and 92.60 gigaflops. Finally, dense connection in Hbase with other modules increases the parameters requirement by 0.01 million and one gigaflops. Our final model has a computational complexity of 93.60 gigaflops and 38.77 million parameters.

3.3 Summary

This contributory chapter presents details of ACNet and HybridNet for SS. ACNet utilized dense connections with multi-context blocks for effective feature representation by encoder units. Attentive residual connections ensure meaningful feature exchange through skip connections. Graph convolution is used to capture the long-range dependencies. The proposed model is trained without any pretrained weights and produces better pixel results than the SOTA methods. Comparative studies for quantitative evaluation and ablation results are also presented.

In HybridNet, densely connected multi-context, multi-resolution models and multiple far-pixel interaction approaches are used to handle short and long-range pixel dependencies

and produce semantically rich object boundaries. The proposed model's dense connections help preserve the original resolution and allow better gradient flow. In addition, a lightweight self-attention module is used for obtaining positional and channel-wise attention, which results in better segmentation maps. A generalized Hough transform-based deep voting module is also used to extract pixel dependencies. Detailed experiments on the ablation study and performance comparison are presented. The proposed model improved 0.7% and 0.2% in F_1 and OA values, respectively, for benchmark evaluation of the $Vaihingen$ dataset. It enhanced the F_1 and OA by 0.81% and 0.2% in local evaluation results for the $Vaihingen$ dataset. For the $Potsdam$ dataset, it improved 0.04% and 0.03%, respectively, in F_1 and OA values from the previous results. The experimental results reveal that the proposed scheme achieved SOTA local and benchmark evaluation results.

In the next contributory chapter, details of proposed models for the Change Detection (CD) task are presented. CD is an important Remote Sensing Vision Task (RSVT) as it provides the temporal interpretation of physical changes happening on the earth's surface.

"Real knowledge is to know the extent of one's ignorance."

~Confucius

4

Multitask learning enhanced aerial change detection

This chapter presents our second contribution to the Change Detection (CD) task. CD for remotely sensed data required two co-paired images for different time intervals. It denotes the physical changes that happened on the earth's surface during that interval. CD is a crucial Remote Sensing Vision Task (RSVT) as it is utilized in monitoring applications that need multi-temporal analysis such as urban sprawl analysis, forest growth/loss assessment, environment change assessment, crop growth analysis, disaster impact assessment, etc. Identifying physical changes on the ground and avoiding spurious changes due to other reasons like co-registration issues, change in illumination conditions, sun angle, and

presence of cloud and fog is a challenging task. In this chapter, two deep models named Difference image Reconstruction enhanced Multiresolution Network (DRMNet) and Triad Multitask Learning for change detection (TMLNet) are proposed, which effectively utilize Multitask Learning (MTL) and produce improved results for CD. The main contributions of this chapter are summarized below:

- The proposed DRMNet achieved efficient long-range dependency mapping and enhanced feature representation with MTL for effective CD.
- To further improve the performance on CD, TMLNet is proposed with triad reconstruction, enhanced backbone, multicontext local self-attention, and MTL.

The rest of the chapter is organized as follows: In Section 4.1 presented details of DRMNet, experiments, and achieved results. Section 4.2 showcased details of TMLNet, experimentation, and results attained by it. This chapter is concluded in Section 4.3.

4.1 DRMNet

Identifying changes in bi-temporal remotely sensed images is very useful in natural resource monitoring, urban planning, land monitoring, and other disaster mitigation applications. Deep Learning (DL) based approaches offered the combination of pixel and object level CD. For DL based methods, two approaches are frequently used, early fusion networks and siamese networks. In an early fusion network [41, 56, 251, 252], input images are first combined and then passed to the network to detect the changes, while in siamese networks [54, 65, 66, 253, 254], the input images are passed to the parallel stream of the network. To model the local context, large-size kernels, local attention, and dilated convolutions are used [40,50,51,56,66,255,256]. Identification of long-range dependency between pixels is required to differentiate between actual and spurious changes, which are achieved using nonlocal operators and self-attention modules [47,53,54,257]. Dense connectivity plays an important role in network design for CD in high-resolution aerial images. Recent network architectures for CD follow dense connections within the network [42, 50]. Current approaches faced

challenges of class imbalance, spurious change identification, and information loss due to a lack of original resolution representation.

This work proposes a multitask learning based CD model where two parallel pipeline architectures predict change map and image difference. The major contributions in DRMNet are as follows:

1. The proposed DRMNet is a multitasking bi-temporal CD model that can efficiently model the long-range pixel dependencies for a very high-resolution aerial image using a multi-scale attention module (MSAM).

2. The DRMNet also incorporates a deconvolution with a sub-pixel convolution module (DSCM) to get more precise change maps by enhancing the quality of feature representation generated by the backbone network (BN).

3. In the DRMNet, a modulus difference-based loss function is used to detect the changes more precisely by the DSCM.

In this work, we have proposed a dense high-resolution network to mitigate the shortcomings mentioned above in the existing literature. Two input images (of which the changes are detected) and the modulus of their difference are concatenated and treated as input in this work. This input is fed to the Backbone Network (BN) to preserve the original resolution and extract high-resolution features. Apart from the BN modules, the proposed model has two other significant modules, named Multi-Scale Attention Module (MSAM) and Deconvolution with Sub-pixel Convolution Module (DSCM). MSAM models the long-range dependencies between pixels by using self-attention maps. DSCM helps to remove the spurious changes effectively to produce a more accurate image difference using a deconvolution layer through sub-pixel convolution. In addition, two different loss functions are used to tackle the class imbalance problem.

Algorithm 1: An algorithm for DRMNet training. Here I_1, I_2 are input images of two different timestamps, $|I_1 - I_2|$ is the modulus of input pair difference, GT is the ground truth, O_1 is the final change map, O_2 is auxiliary output, and L_1, L_2 are calculated loss values.

Data: $Sample_{train} = (I_1, I_2, GT) \in Trainset$
Result: O_1
while $exist(Sample_{train})$ **do**
  $I \leftarrow concat(I_1, I_2, |I_1 - I_2|)$;
  $F \leftarrow BN(I)$;
  $O_1 \leftarrow MSAM(F)$;
  $F_{downsample} \leftarrow Downsampling(F)$;
  $O_2 \leftarrow DSCM(F_{downsample})$;
  $L_1 \leftarrow Loss_1(O_1, GT)$;
  $L_2 \leftarrow Loss_2(O_2, |I_1 - I_2|)$;

4.1.1 Proposed Network Architecture

The proposed DRMNet model has three primary modules, as depicted in Fig. 4.12. The first one is Backbone Network (BN), which is an extension of an existing model named HRNet [15]. The second is the Multi-Scale Attention Module (MSAM) to generate self-attention at different scales. The third one is Deconvolution with Sub-pixel Convolution Module (DSCM), which is used to predict the difference (of two input images) reconstruction. The proposed DRMNet model takes two images (i.e. image A and image B, where we have to detect changes between A and B) and the corresponding image difference ($|A-B|$) as input. First, the concatenation of three images (A, B, and $|A - B|$) is passed through a sequence of Convolution, Batch Normalization, and ReLU layers, and then its output is fed into BN. This module outputs 48 channels feature matrices of dimension $N \times N \times 48$ where $N \times N$ is the input image(s) dimension. We selected 48 channels output after initial layers based on the ablation study of different initial channel configurations as presented in Table 4.6. The output is fed to the MSAM module to identify whether the image is changed or not, and the inference is tallied with ground truth to find the corresponding loss. The exact output from the BN is down-sampled, and the down-sampled version is fed into the DSCM. The output of the DSCM is compared with the input image differences to generate the corresponding loss (L2).

4.1.1.1 Backbone Network (BN) module

This module extracts the high-resolution image features while preserving the original image resolution. The block diagram of the Backbone Network (BN) architecture is depicted in Fig. 4.1. We used residual block as the basic unit for this backbone. Residual connections can be represented as

$$R(x) = \kappa(x) + x \tag{4.1}$$

Here, x is input to block with residual connection, κ is nonlinear mapping, and R is residual block output. We extended an existing network called HRNet proposed by Wang et al. [15] where the high-resolution representations of the input image are maintained by connecting the high-to-low resolution convolution streams in parallel and by exchanging the information across resolutions frequently. Likewise, in this work, we generated multiple streams of different resolutions (e.g., 1, $\frac{1}{2}$, $\frac{1}{4}$, etc.) from the input image. These multi-resolution streams are fully connected, as shown in Fig. 4.1. Features of different resolutions are combined using the addition operator. We extended the HRNet model by adding the skip connections between the same resolution stream for parameter sharing and better gradient flow. This Backbone Network (BN) module provides high-resolution feature extraction and multi-resolution information fusion, which are required to identify actual changes in high-resolution images.

4.1.1.2 Multi-Scale Attention Module (MSAM)

As described in Fig. 4.12, the output of the Backbone Network (BN) is fed into two different modules in parallel. One is the Multi-Scale Attention Module (MSAM), and the other is the Deconvolution with Sub-pixel Convolution Module (DSCM). In this sub-section, MSAM is described with an illustrative example as given in Fig. 4.2. The self-attention [258] is used at multiple lower scales to produce aggregated multi-scale self-attentive features. This module essentially generates self-attention maps at different lower scales such as $[\frac{1}{4}, \frac{1}{6}, \frac{1}{8}]$ of the original resolution and fuses them as described in the Fig. 4.2. The input (original resolution) features are down-sampled in different lower resolutions $(\frac{1}{4}, \frac{1}{6}, \frac{1}{8})$ and passes

through the Self Attention Module (SAM) individually. Then all three SAM responses are up-sampled respective to their original resolutions and then added to get the final response. The self-attention is used to find the long-range dependencies between pixels. The block diagram of SAM is depicted in Fig. 4.3. Self-attention captures the global interaction of the pixels using dot product operation between the linear representation of inputs. For the given input I, it can be represented as

$$Z = f(\theta(I), \psi(I)) * \mu(I) + I \tag{4.2}$$

where Z is self attentive output map, f is mapping function used in self attention, $\theta(.), \psi(.), \mu(.)$ are linear functions. The linear functions are implemented using 1×1 convolutions. These linear functions produced a linear transformation of input I which is utilized by mapping function f and distributed over linear representation generated by $\mu(.)$. The function f uses multiplication and softmax operation and is defined as

$$f(\theta(I), \psi(I)) = softmax(\theta(I) * \psi(I)) \tag{4.3}$$

Creating a self-attention map in the original resolution is very memory extensive and not feasible for large image patches. In this module, we propose to generate a self-attention map at multiple lower resolutions and later combine the up-sampled maps. This approach produced better results than self-attention maps at a single resolution and required fewer computational resources than self-attention at the original resolution. The output of MSAM (Z_{final}) can be represented by the following equations.

$$Z_{final} = Up_{4x}(Z_{\frac{1}{4}}) + Up_{6x}(Z_{\frac{1}{6}}) + Up_{8x}(Z_{\frac{1}{8}}) \tag{4.4}$$

Where $Z_{\frac{1}{4}}, Z_{\frac{1}{6}}, Z_{\frac{1}{8}}$ are self attentive maps, generated at $\frac{1}{4}^{th}, \frac{1}{6}^{th}, \frac{1}{8}^{th}$ resolution of input I. $Up_{4x}, Up_{6x}, Up_{8x}$ represent upsampling operations with upsampling rates of 4, 6, and 8. It is used to produce the aggregated attentive map (Z_{final}) at the original resolution.

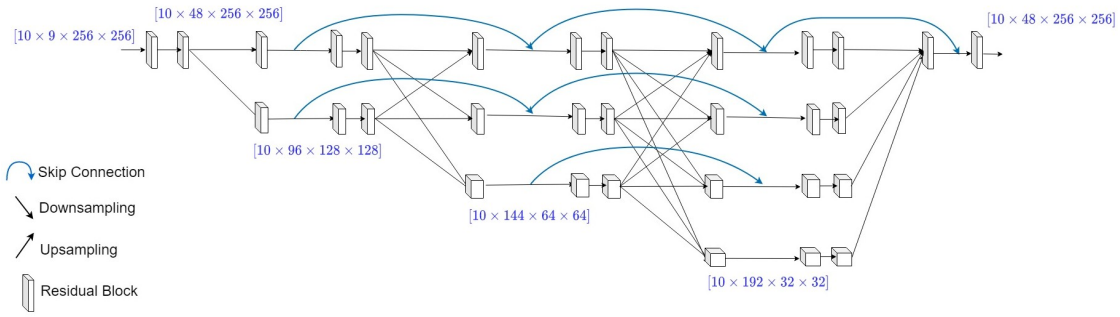


Figure 4.1: *Backbone module. Here, blue connections represent proposed skip connections between the same resolution stream. The down arrow represents downsampling, and the up arrow represents upsampling of features.*

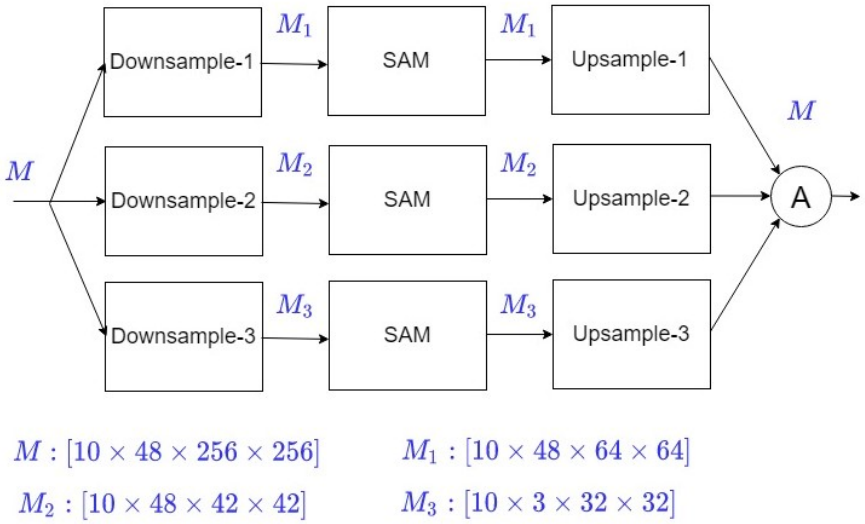


Figure 4.2: *Multi-Scale Attention Module. Here, SAM is the self-attention module. Downsample-1, Downsample-2, Downsample-3, and Upsample-1, Upsample-2, Upsample-3 represent downsampling and upsampling operations at different scales.*

4.1.1.3 Deconvolution with Sub-pixel Convolution Module (DSCM)

This module is a deconvolution (decoder) module, which takes a sub-sampled version of image difference features (generated by the BN module) and reconstructs the image difference in the original resolution. The resolution scaling (upsampling) is made through the sub-pixel convolution process. It has been experimentally observed that by predicting image differences from down-sampled feature maps, this module can remove the spurious changes effectively to produce a more accurate image difference. Thus, the DSCM is one of the novel contributions of the proposed work. A basic block diagram of this DSCM is depicted

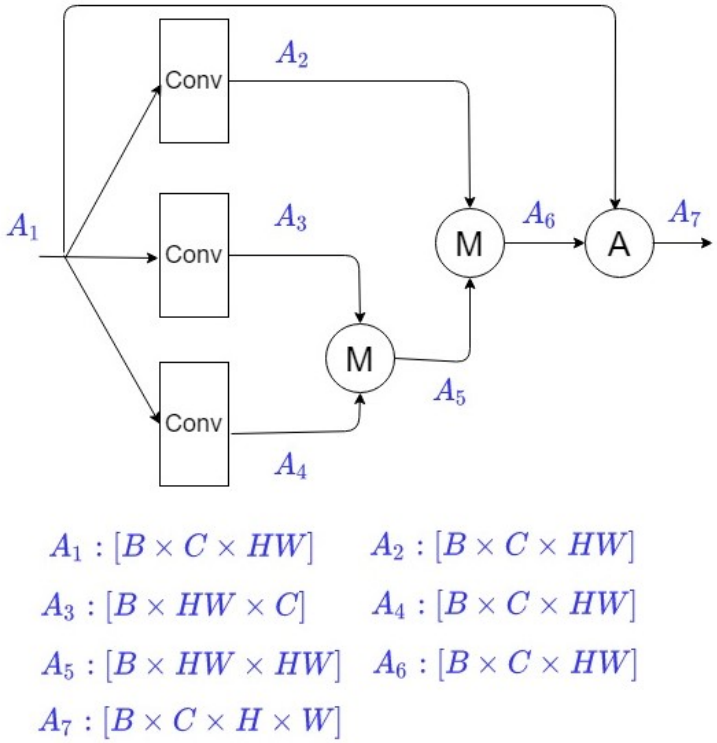


Figure 4.3: *Self Attention Module, M represents multiplication operation, and A is addition operations.*

in Fig. 4.4. The DSCM consists of two convolutions (with Tanh activation function) layers, one convolution layer with subpixel convolution, and a sigmoid layer. The sub-pixel convolution [259] is defined as a standard convolution in low-resolution space followed by a periodic shuffling operation. It is used as a part of the proposed deconvolution process for the required up-sample process to reconstruct the image difference in the original resolution. Architecturally, this pipeline (BN module, down-sampler and followed by DSC module as deconvolution process) forms an encoding-decoding like architecture which helps to remove the spurious changes (due to image acquisition, spatial de-synchronization, etc.) from the image difference (i.e., the change map). Sub-pixel convolution is a well-known technique that transforms the input of size $H \times W \times C$ to $(H \times d) \times (W \times d) \times \frac{C}{d^2}$ with H, W, C are height, width, and channels of input and d is the upsampling factor. We have used the sigmoid activation function because it gives output in the range of [0, 1]. This module can be mathematically formulated as

$$O_2 = \sigma(PS(w_3(\tanh(w_2(\tanh(w_1(O_{F_{\frac{1}{2}}}) + b_1) + b_2) + b_3)) \tag{4.5}$$

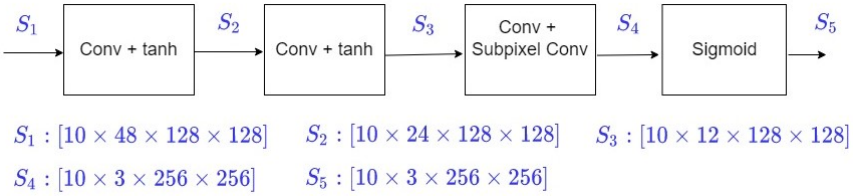


Figure 4.4: *Deconvolution with Sub-pixel Convolution Module (DSCM). Here, Conv is the convolutional layer.*

Here, PS is pixel shuffle convolution, $O_{F_{\frac{1}{2}}}$ is downsampled output of backbone, w_i is weigh of convolution kernel and b_i is bias of convolution.

The output of this DSCM is compared to the modulus of input pairs difference using the loss function (L_2), which is calculated as means square error (MSE). This loss function trains the BN module for better feature generation. The features generated from the BN module are downsampled and fed to DSCM to predict the change map by removing the spurious contents from the input features. To get the loss, these predicted change maps (O_2) are compared with the image differences (using the L_2 loss function).

4.1.2 Loss function

In Fig. 4.12, it can be observed that two loss functions are used in the proposed model. First loss function (L_1) is calculated between O_1 and ground truth (G). In this case, we have used a combined loss function as it has been used in SNUNet [50]. The authors have argued that there exists a sample imbalance effect as the number of unchanged pixels is often far more than the number of changed pixels. To reduce the sample imbalance effect, a combination of weighted entropy loss L_{ce} and dice loss L_d has been used. L_{ce} and L_d can be represented as per the following equations:

$$L_1 = L_{ce} + L_d \tag{4.6}$$

$$L_{ce} = -\sum_{1}^{N} c \times \log(S(m)_c) \tag{4.7}$$

$$L_d = 1 - \frac{2 \times O \times S(m)}{O + S(m)} \tag{4.8}$$

Here, $S(.)$ is the softmax function, c is a class vector, m is the model's output before the softmax layer, and O is the ground truth.

The second loss function (L_2) is calculated as means square error (MSE) between O_2 and $|I_1 - I_2|$. The overall loss function is calculated as follows

$$Loss(I_1, I_2, G, O_1, O_2) = L_1(G, O_1) + \alpha \times L_2(|I_1 - I_2|, O_2) \tag{4.9}$$

Here, L_{ce}, L_d, α and L_{mse} are weighted cross-entropy loss, dice loss, loss weight constant (variable with a value between [0, 1]) and mean square error loss. L_1 is calculated with the predicted change map using the output of MSAM (O_1) with the actual ground truth, and L_2 loss is calculated between the predicted image difference using the output of DSCM (O_2) and the ground truth. We find the optimal value of α using exhaustive experiments as listed in Table 4.5. From this, we found that the α value 0.9 produces the optimal performance.

4.1.3 Datasets

We have used the Google dataset CDD proposed by Lebdev et. al. [260], LEVIR-CD dataset introduced by Chen et. al. [53], WHU BCDD [261] dataset, and SYSU [49] dataset. CDD dataset consists of 16000 images of 256 × 256 pixels containing RGB data with training, validation, and test sets sizes of 10000, 3000, and 3000, respectively. LEVIR-CD dataset contains 637 bitemporal images of size 1024 × 1024 pixels with RGB bands. We have generated 256 × 256 size patches from the LEVIR-CD dataset for our experiments. WHU BCDD dataset consists of a training image of size 15354 × 21243 pixels and a testing image of size 15354 × 11265 pixels. Earlier, a single image of size 15354 × 32507 was provided, and the test image was generated randomly. We cropped 256 × 256 size patches from the training image to train the model. SYSU dataset consists of 20000 aerial images for CD with the train set, validation set, and test set divided into 12000, 4000, and 4000 images.

CDD dataset contains changes of buildings, roads, etc. SYSU dataset contains changes

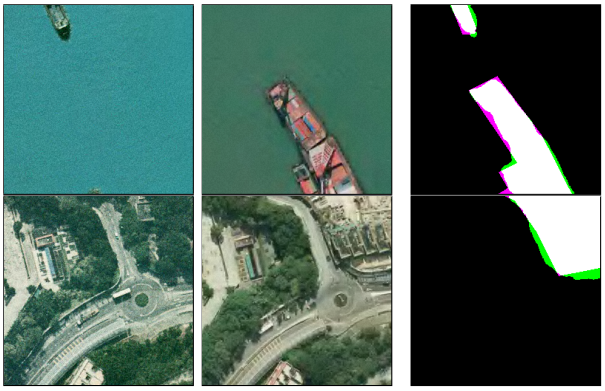


Figure 4.5: *Output visualization of* 256×256 *pixels patches for SYSU test dataset. Here, the white color is the actual change detected by the model, green color is psuedo change detected by the model, and the pink color is the actual change missed.*

of building, road, sea construction, vegetation, construction, etc. LEVIR-CD and WHU BCDD datasets focus on building CD.

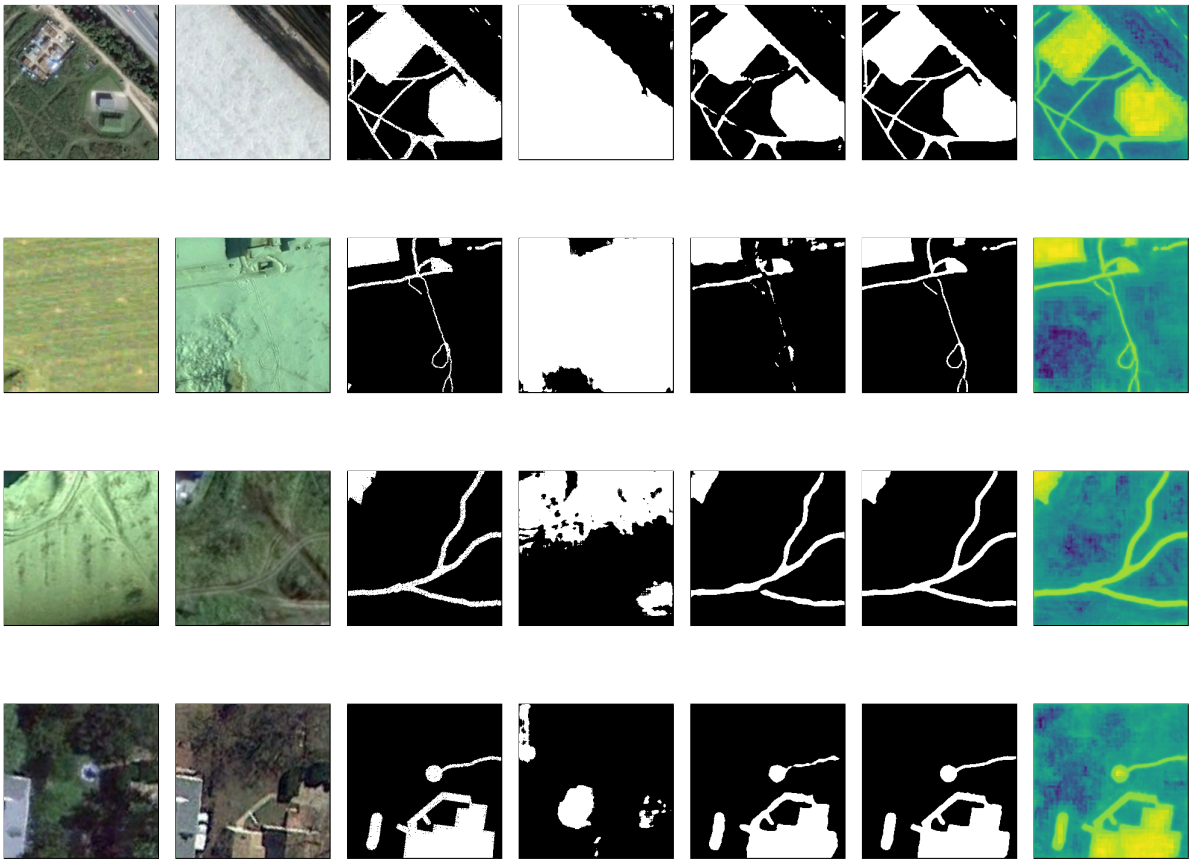
4.1.4 Experiments

4.1.5 Training and Hyperparameters

We have used an initial learning rate of 0.001 and a batch size of 10 for the training of our network. We used the Nvidia P100 graphics card with 32 GB of graphics memory. We selected batch size based on this available graphics memory. We used data augmentations of horizontal and vertical flipping with 50% probability and random rotation to increase their variability. For each dataset, the model is trained for 300 epochs with early stopping when it cannot optimize further. Input pairs and ground truths are normalized before passing to the network. During inference, we used test time augmentation of horizontal and vertical flipping and 90° rotations. Final maps are produced using the average of all augmented outputs.

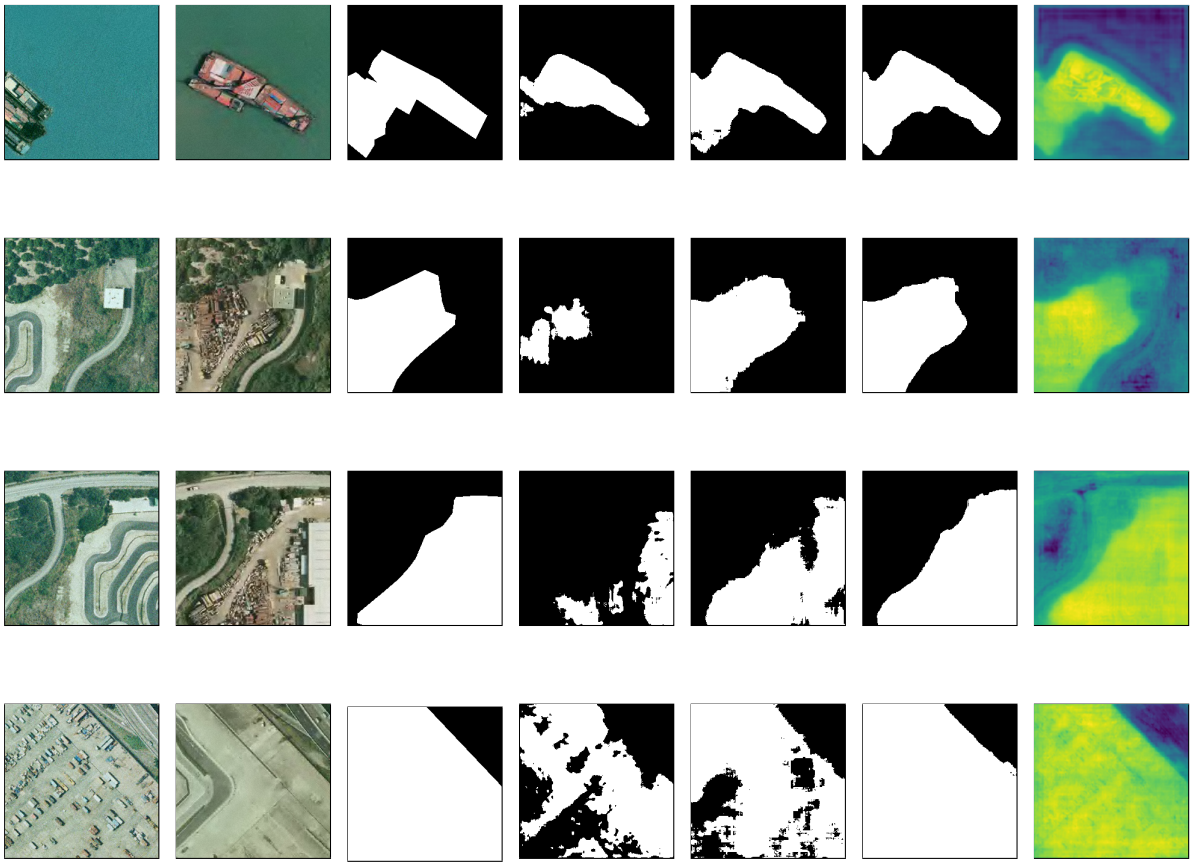
4.1.6 Comparative Analysis

We used Precision (P_r), Recall (R_c), F1 Score (F_1), Intersection over Union (IoU) and Overall Accuracy (OA) as performance metrics for quantitative comparison of outputs. These



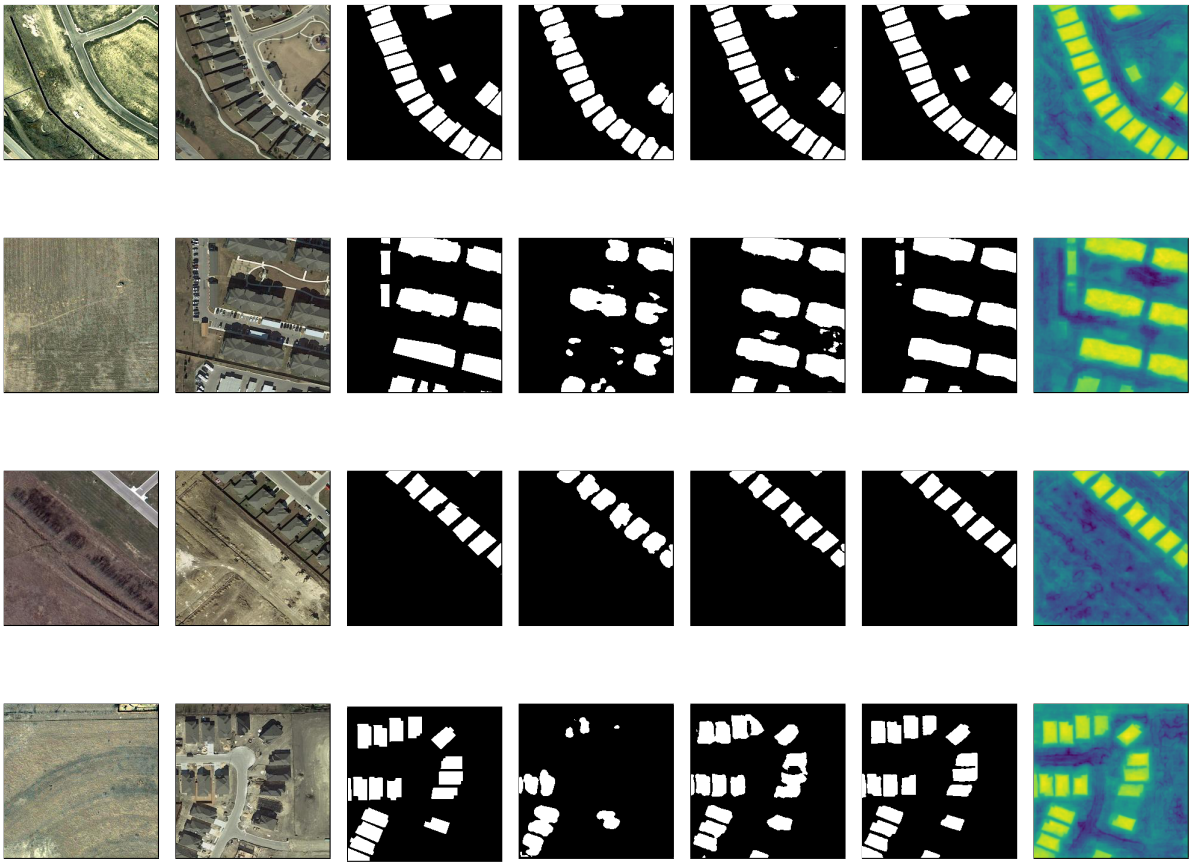
(a) *Image 1* **(b)** *Image 2* **(c)** *GT* **(d)** *FC-EF* **(e)** *SNUNet* **(f)** *Ours* **(g)** *Heat map*

Figure 4.6: 256×256 *pixels patches output visualization for CDD test dataset.*



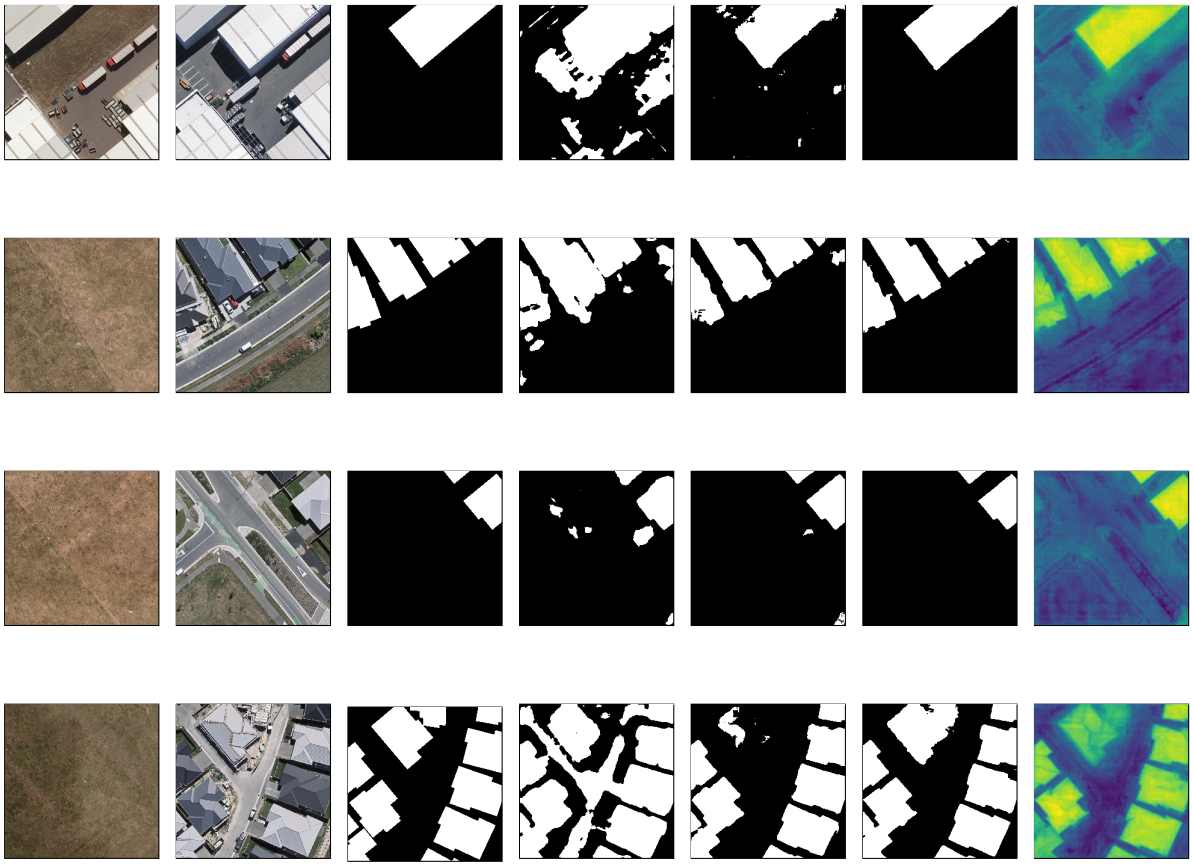
(a) *Image 1* **(b)** *Image 2* **(c)** *Ground truth* **(d)** *FC-EF* **(e)** *SNUNet* **(f)** *Our* **(g)** *Heat map*

Figure 4.7: 256×256 *pixels patches output visualization for SYSU test dataset.*



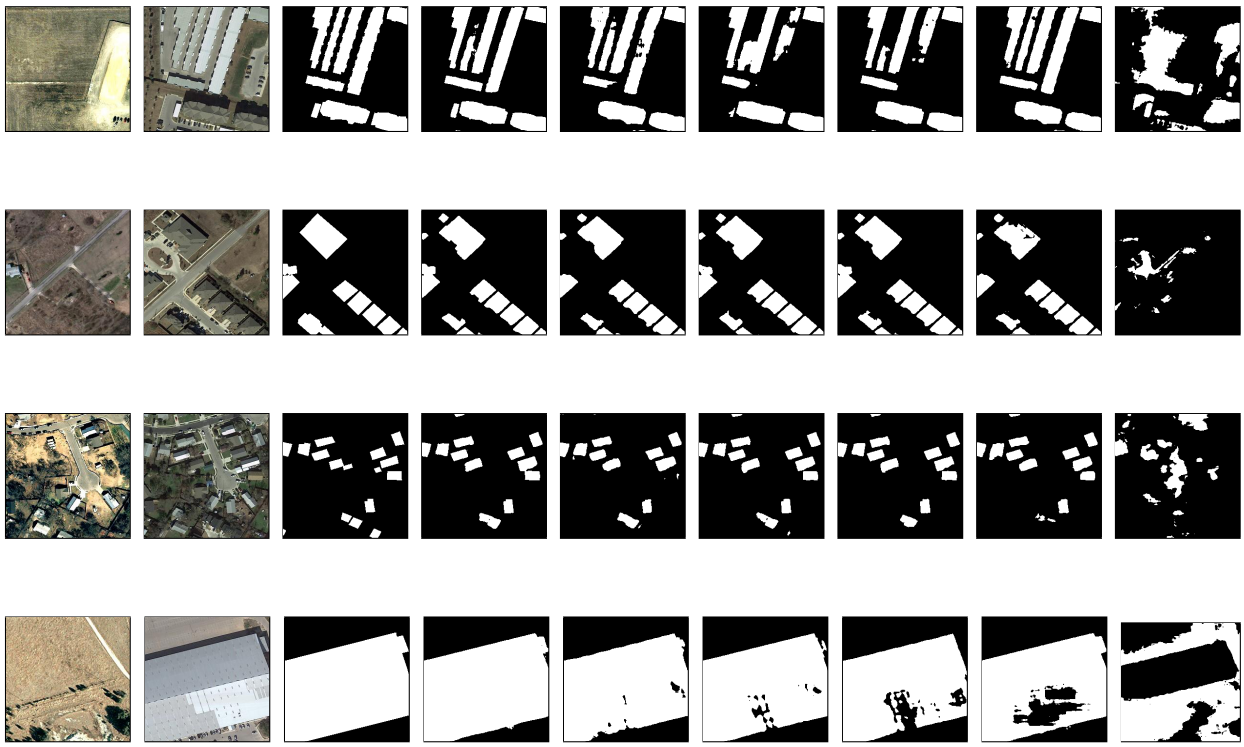
(a) *Image 2* (b) *Image 2* (c) *Ground truth* (d) *FC-EF* (e) *SNUNet* (f) *Ours* (g) *Heat map*

Figure 4.8: 256×256 *pixels patches output visualization for LEVIR test dataset.*



(a) *Image 1* (b) *Image 2* (c) *Ground truth* (d) *FC-EF* (e) *SNUNet* (f) *Ours* (g) *Heat map*

Figure 4.9: 256×256 *pixels patches output visualization for WHU test dataset.*



(a) *Image 1* (b) *Image 2* (c) *GT* (d) *BMDRDa* (e) *BMDR* (f) *BMD* (g) *BM* (h) *BS* (i) *B*

Figure 4.10: *Ablation study visualization with* 256×256 *pixels patches. Here, B is Base network, D is DSCM, S is SAM, M is MSAM, R is residual connections, Da is data augmentation at test time.*

(a) *Image 1* **(b)** *Image 2* **(c)** *Ground truth* **(d)** *Our Output*

Figure 4.11: 1024×1024 *pixels output visualization for LEVIR test images.*

metrics are used in recently published works for comparative study. These are calculated using true positive (t_p), false positive (f_p), true negative (t_n) and false negative (f_n) values as per following equations:

$$P_r = \frac{t_p}{t_p + f_p} \tag{4.10}$$

$$R_c = \frac{t_p}{t_p + f_n} \tag{4.11}$$

$$F_1 = 2 \times \frac{P_r \times R_c}{P_r + R_c} \tag{4.12}$$

$$OA = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \tag{4.13}$$

$$IoU = \frac{t_p}{t_p + f_p + f_n} \tag{4.14}$$

Numerical comparisons against the SOTA methods on different datasets are presented in Table 4.1, 4.2, 4.3 and 4.4 with the best result highlighted in the red color. As it can be observed in Table 4.1, for the CDD dataset, our model achieved 99.57 % OA with *F1 Score* of 98.20 which is 0.44% and 1.66 % respectively higher than the SOTA result [56]. The corresponding subjective (visual) results are presented in Fig. 4.6. As per Table 4.2

and Fig. 4.8, it is observed that our model achieved *F1 Score* of 91.97% which is 0.14% higher than the SOTA result [257] for LEVIR-CD dataset. From Table 4.3 and Fig. 4.7, it is observed that against the SYSU dataset, we achieved 91.23% OA and 80.53% *F1 Score* which are 1.27% and 1.61% higher respectively than the SOTA result [49]. For the BCDD dataset, we reported our result on a separate test image provided in the dataset. Earlier approaches have reported results on the random test split. These results are not comparable with our output. For comparison against the BCDD dataset, we have trained SNUNet [50] and generated results on the test image. Our results are presented in Table 4.4 where not comparable results are shown with * mark and visual comparison shown in Fig 4.9.

In Table 4.5, detailed experiment results are presented for finding the optimal value of α. We trained our final model on the LEVIR-CD dataset. The experiment started with an initial value of 0.5 for α and achieved $F1Score$ of 91.09% and IoU of 83.56%. We repeated this experiment with an increase of 0.1 in the α value till the final value of 1.0. After analyzing the result produced, we found that 0.9 is the optimal value.

4.1.7 Visual Analysis

Visual test results are shown in Fig. 4.5 using RGB composite (ground truth in red and blue band, prediction in green band) images to demonstrate the accurate prediction capability of the proposed model. We can conclude from these figures that our model produced results very similar to actual ground truth. It missed a small number of actual changes with few detection of pseudo changes. The visual output comparison of proposed work, FC-EF [251], SNUNet [50], and groundtruth is presented in Fig 4.18-4.23. The competitive methods are chosen based on their recent SOTA performance. For example, SNUNet [50] is the recently published one of the SOTA works. FC-EF [251] is chosen as it is one of the foundation works for DL based CD. It is [251] used as a baseline to show how the proposed model improved the performance over the base performance. In Fig. 4.18, CDD dataset is used with changes of buildings and roads. While FC-EF [251] network could not make a difference between building and road changes, SNUNet [50] generated better results. Our network produced

more precise results compared to these. Fig. 4.22 contains outputs produced on the SYSU dataset with multiple change types. For this dataset, the proposed method is relatively better than others. The output comparison on the LEVIR-CD dataset with building change instances is shown in Fig. 4.19. The output of FC-EF [251] missed many building change instances and SNUNet [50] detected some false changes. Our proposed work output is close to the actual ground truth and missed minimal changes. We used another building change dataset, WHU-CD, and its visual outputs are shown in Fig. 4.23. For this dataset, FC-EF [251] detected several false changes, and SNUNet [50] also detected some false changes. Our proposed work detected fewer false changes.

	Precision	Recall	F1 Score	IoU	OA
ADS-Net [40]	89.79	79.58	82.72		
UNet++ [252]	89.54	87.11	87.56		
IFN [256]	94.96	86.08	90.30		97.71
BA^{2}Net [255]	88.12	95.28	91.36		98.94
BSFNet [66]	90.5	93.3	91.9		98.10
DASNet [54]	93.2	92.2	92.7		98.2
SRCDNet [62]	-	-	92.94		
DiffUNet++ [42]	92.15	94.63	93.37		
DSAMNet [49]	94.54	92.77	93.69	88.13	
DDCNN [41]	96.71	92.32	94.46	89.51	98.64
SNUNet [50]	96.3	96.2	96.2		
LSS-Net [262]	96.74	95.87	96.30	-	-
AGCDetNet [56]	95.03	98.10	96.54	-	99.13
Ours	97.92	98.49	98.20	96.46	99.57

Table 4.1: *CDD dataset performance comparison*

4.1.8 Ablation Study

We conducted extensive experiments to find the optimal fusion strategy for the inputs, the optimal number of initial channels, the number of resolution levels in MSAM, and the architecture components. We conducted all experiments on the LEVIR-CD test dataset.

	Precision	Recall	F1 Score	IoU	OA
BSFNet [66]	82.7	94.0	88.0	-	97.0
STANet [53]	83.8	91.0	87.3		
CDNet + IAug [58]	91.6	86.5	89.0	-	-
DiffUNet++ [42]	92.4	87.1	89.6	-	-
ADS-Net [40]	89.67	91.36	89.80	-	-
BiT [47]	89.24	89.37	89.31	80.68	98.92
SNUNet [50]	90.61	89.01	89.80	81.49	98.97
DDCNN [41]	91.85	88.69	90.24	82.21	98.11
AGCDetNet [56]	92.12	89.45	90.76	83.09	-
CEECNet [257]	93.81	89.92	91.83	84.89	-
Ours	93.05	90.91	91.97	85.13	99.19

Table 4.2: *LEVIR-CD dataset performance comparison*

	Precision	Recall	F1 Score	IoU	OA
FC-EF [251]	74.32	75.84	75.07	60.09	
BiDateNet [253]	81.84	72.60	76.94	62.52	
STANet [53]	70.76	85.33	77.37	63.09	
DSAMNet [49]	74.81	81.86	78.18	64.18	
SNUNet [50]	78.16	79.68	78.92	65.18	89.96
Ours	84.55	76.86	80.53	67.39	91.23

Table 4.3: *SYSU dataset performance comparison*

4.1.8.1 Ablation for fusion strategy

We started with a base model similar to HRNet [15]. It is observed in the literature that handling multi-temporal inputs is an important task for CD. For an optimum strategy to combine the input pair, a comprehensive set of experiments has been conducted as tabulated in Table 4.7. We tried an early fusion of input pairs, a medium fusion of input pairs, no fusion of input pairs, and early fusion with the modulus of the input pair difference. The input pair is concatenated in early fusion and passed to the network for feature extraction. In medium fusion, the input pair is given simultaneously to the base model till the middle of it. After that, features from both streams are combined using a concatenation operation. No fusion strategy did not use concatenated input features till the final class convolution layer. In Early fusion with the modulus of the input pair difference approach, concatenated features of input pairs and the modulus of input pair difference are used. The experiment shows that early fusion with the difference approach gives the best results.

	Precision	Recall	F1 Score	IoU	OA
*BiT [47]	86.64	81.48	83.98	72.39	98.75
*DDCNN [41]	91.85	88.69	90.24	82.21	98.11
*AGCDetNet [56]	92.12	89.45	90.76	83.09	-
*CEECNet [257]	95.57	92.043	93.77	88.23	-
*FCCDN [263]	96.39	91.24	93.73	88.20	-
*DTCDSCN [51]	-	89.32	89.01	78.08	-
*LSS-Net [262]	94.18	93.36	93.77	-	-
SNUNet [50]	85.25	81.09	83.13	71.12	98.80
Ours	87.93	84.37	86.11	75.61	99.01

Table 4.4: *WHU BCDD dataset performance comparison, Here, * represents results computed on different split for test data and not comparable*

α	0.5	0.6	0.7	0.8	**0.9**	1
F1 Score	91.09	91.15	91.18	91.29	**91.46**	91.43
IoU	83.56	83.77	83.89	84.19	**84.28**	84.25

Table 4.5: *Ablation study for loss parameter value*

4.1.8.2 Ablation for initial channels

The selection of an optimal value for the initial number of channels of BN is essential as it affects the computational requirement of the model. We find the number of initial channels for the base model through detailed experiments as shown in Table 4.6. We tried three different combinations of 32, 48, and 64 as initial channels. From the experiment, it is found that the initial channel of 48 achieves the best results as it produces *F1 Score* similar to the 64 channels but with lesser computational cost.

4.1.8.3 Ablation for attention mechanism

In Table 4.8, we presented *F1 Score* produced by attention modules with different resolution inputs. We used an input size of $1 \times 48 \times 256 \times 256$ for the calculation of the computational complexity of attention modules. We started with SAM with input of $\frac{1}{4}^{th}$ of the original resolution and extended it to a multi-scale self-attention module. To find the optimal number of different resolution inputs to be used we started with MSAM-2 where input resolutions of $\frac{1}{4}^{th}$ and $\frac{1}{6}^{th}$ are utilized. In MSAM-3, input resolutions of $\frac{1}{4}^{th}$, $\frac{1}{6}^{th}$ and $\frac{1}{8}^{th}$ are used. MSAM-4 took $\frac{1}{16}^{th}$ resolution input additionally. From the experiments, we have found that MSAM

	Parameters	Precision	Recall	F1 Score
Base-32	15.67 M	90.01	86.99	89.11
Base-48	34.94 M	90.67	88.79	89.72
Base-64	61.70 M	90.70	88.78	89.73

Table 4.6: *Ablation study for number of initial channels to be used. Parameters are calculated in millions (M)*

	Precision	Recall	F1 Score
No fusion	89.86	77.90	83.46
Medium	90.99	85.30	88.05
Early	89.38	87.97	88.67
Early + difference	90.67	88.79	89.72

Table 4.7: *Ablation study on fusion methods*

with three resolution features is giving better results quantitatively and with respect to the computational complexity.

4.1.8.4 Ablation for architecture component

This study is conducted to showcase the importance of each component in the proposed work. The visual comparison of the ablation study for various components of the proposed model is shown in Fig. 4.10. The rest of the ablation experiments (as shown in Table 4.9) are done with the base model with the initial optimum channel number obtained by Table 4.6 and early fusion with the difference strategy. This model achieved *F1 Score* of 89.72%. After adding the self-attention module (SAM), it increased by 0.47%. We improved the SAM to MSAM with the base model, and this increases the *F1 Score* to 90.38 % and by 0.66%. We added residual connections in the base model, and this increases the *F1 Score* to 90.04% and by 0.32% from the base model. To further improve the performance, we have added the DSCM. First, we combined the DSCM and MSAM with the base and achieved *F1 score* of 90.61% which is 0.89% higher than the base model. Further inclusion of skip connection has improved it to 91.46% which is 1.74% higher than the base model. After applying test time augmentation, this further improved to 91.77 %. In addition, overlapping strides are used in output as utilized in [257]. Output strides of 128 pixels resulted in the *F1 Score* of 91.88, and with stride 64, it reached 91.97%. This study shows that skip connection in

	Parameters	FLOPS	F1 Score
SAM	9.6 K	44.37 M	90.19
MSAM-2	10.17 K	45.32 M	90.25
MSAM-3	10.75 K	45.83 M	90.38
MSAM-4	11.33 K	46.22 M	90.38

Table 4.8: *Ablation study on attention module. Here, SAM is a self-attention module. MSAM-2 is the multi-scale attention module with two resolutions of self-attention maps. MSAM-3 is a multi-scale attention module with three resolutions of self-attention maps. MSAM-4 is a multi-scale attention module with four resolutions of self-attention maps. Parameters are calculated in thousands (K), and FLOPS are counted in millions (M).*

	Precision	Recall	F1 Score	IoU
Base	90.67	88.79	89.72	80.94
Base + Residual connection	90.31	89.79	90.04	81.03
Base +SAM	90.59	89.81	90.19	82.01
Base + MSAM	90.83	89.95	90.38	82.12
Base + DSCM + MSAM	89.67	91.58	90.61	82.21
Base + DSCM + MSAM + Residual connection	92.09	90.85	91.46	84.28
Base + DSCM + MSAM + Residual connection + DA	92.78	90.77	91.77	84.79
Base + DSCM + MSAM + Residual connection + DAE1	92.96	90.83	91.88	84.98
Base + DAE2	90.69	88.82	89.75	80.94
Base + Residual connection + DAE2	90.36	89.81	90.08	81.04
Base + SAM + DAE2	90.59	89.83	90.21	82.01
Base + MSAM + DAE2	90.81	90.07	90.44	82.16
Base + DSCM + MSAM + DAE2	89.81	91.74	90.76	82.43
Base + DSCM + MSAM + Residual connection + DAE2	93.05	90.91	91.97	85.13

Table 4.9: *Ablation Study. Here DA is data augmentation at test time, DAE1 is data augmentation at test time with an overlapped evaluation with stride 128 pixels, and DAE1 is data augmentation at test time with an overlapped evaluation with a stride of 64 pixels.*

the base model improved the model's performance. The combination of DSCM and MSAM also significantly enhances the base model's result.

4.2 TMLNet

The difference in features approximately represented the changes, and some works utilized the feature differences [39–48] in the neural network. Further, this approach is expanded for the use of multi-scale feature differences [43, 44, 49]. Focusing on the relevant area of the input is important as it filters out the irrelevant regions. The attention-based [39–42, 48–54, 56, 57, 66] methods do this using attentions maps generated on local and global levels.

Scale plays an important role in the identification of changes. Few methods [40, 42, 45] calculated change maps at multiple scales. Few methods [39, 56, 61] are based on increasing kernel size. Dilated convolutions based networks are proposed to take advantage of multiple context features. Multitasking based methods [48, 51] are proposed where additional tasks are learned with the CD backbone. The CD methods suffer from overfitting problems due to a considerable class imbalance between change and unchanged classes. Effective change localization is also difficult. It is observed that an accurate CD task requires effective feature learning, which needs a noiseless representation. Moreover, CD improves with error-free image reconstruction as an auxiliary task, which needs joint feature representation by a feature extractor.

In this work, we proposed a novel triad (which is a combination of input images and its differences) learning based multiresolution architecture for effective CD. The major contributions are as follows:

1. A triad reconstruction enabling multiscale feature learning is proposed for multiobjective-based joint learning.

2. A multiresolution and local attention based enhanced backbone module (EBM) is designed as the backbone of the proposed model for finer feature extraction.

3. A multi-context local self-attention module (MCLSAM) is proposed to find far pixel dependency with reduced resource requirement.

4. The Laplacian pyramid pooling based loss function is used for feature reconstruction based CD.

The effectiveness of the proposed approach is validated through detailed quantitative comparisons. We achieved SOTA results for seven benchmark datasets.

4.2.1 Proposed Network Architecture

In the proposed triad learning based network architecture named TMLNet, the triad input is passed to the Enhanced Backbone Module (EBM) to extract multi-scale features, which

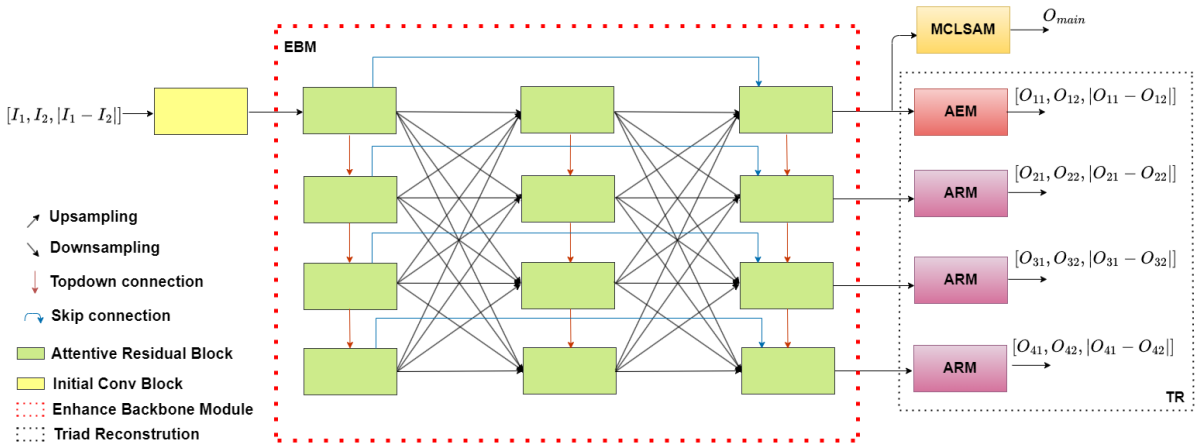


Figure 4.12: *The complete details of TMLNet. Here, MCLSAM is the multi-context local self-attention module, AEM is the autoencoder module, ARM is the attentive reconstruction module, I_1 and I_2 are inputs of the TMLNet, O_{main} is the main output representing changes, O_{i1}, O_{i2} are reconstruction outputs with i resolution level.*

is used for the generation of change maps and triad reconstruction. Extracted original resolution features are utilized in the Multi Context Local Self Attention Module (MCLSAM) for change map creation and in the Auto Encoder Module (AEM) for triad reconstruction in original resolution. Features extracted in low resolutions are used in multiple Attentive Reconstruction Module (ARM) for triad reconstruction at multiple levels as shown in Fig. 4.12. EBM consists of proposed top-down connections with Attentive Residual Blocks for better feature extraction. MCLSAM finds global pixel relation efficiently using proposed multi-context local self-attention. ARM and AEM are used for the proposed MTL based triad reconstruction to improve feature representation capability and to avoid overfitting. A composite loss function is also proposed which consists of CD loss and multi-scale triad reconstruction loss using the Laplacian pyramid. The more details on the proposed components are as follows:

4.2.1.1 Triad Learning

A triad is defined as the combination of the input pair and its difference. For the input image of I_1, I_2 the triad T_I is defined as

$$T_I = concat(I_1, I_2, |I_1 - I_2|) \tag{4.15}$$

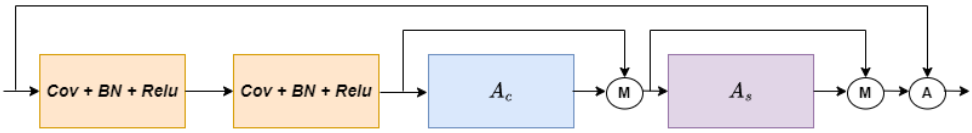


Figure 4.13: *Attentive Residual Block. Here, Conv + BN + Relu is the convolution followed by batch normalization and Relu operations. A_c is channel attention, A_s is spatial attention, M is multiplication operation, and A is addition operation.*

Here, *concat* is the concatenation operation. Triad learning is about feature reconstruction of the triad, and it allows MTL with hard parameter sharing. It also avoids the overfitting of the network [264]. We used $Early + difference$ as a fusion strategy for input pairs as proposed in [48] because it gives comparative better results. triad reconstruction does not significantly increase the computational cost of the network and enables the learning of multiple tasks from the common backbone features.

4.2.1.2 Enhanced Backbone Module

We proposed Enhanced Backbone Module (EBM) as the backbone, which is the enhancement of high-resolution blocks of HRNet [15] and the backbone network of DRMNet [48]. We utilized spatial and channel-wise attention in each residual block of the backbone. This enhances the feature representation capability of EBM and helps in better discrimination of change features. The top-down connection in each stage is proposed as this resulted in better sharing of multi-scale feature information. A dense connection is added in multiple stages of the backbone for enhanced weight sharing with the same scale. The channel and spatial attention are calculated using the following equations.

$$A_c = \sigma(conv((SP_{avg}(F_{in}) + SP_{max}(F_{in}))) \tag{4.16}$$

$$A_s = \sigma(conv(concat((CP_{avg}(F_{in}), CP_{max}(F_{in})))) \tag{4.17}$$

Here A_c is channel attention map, A_s is spatial attention map, *conv* is convolution operation, SP is spatial pooling, CP is channel-wise pooling, and F_{in} is feature inputs. These A_c and A_s are combined in the residual block as shown in Fig. 4.13. It can be depicted as the

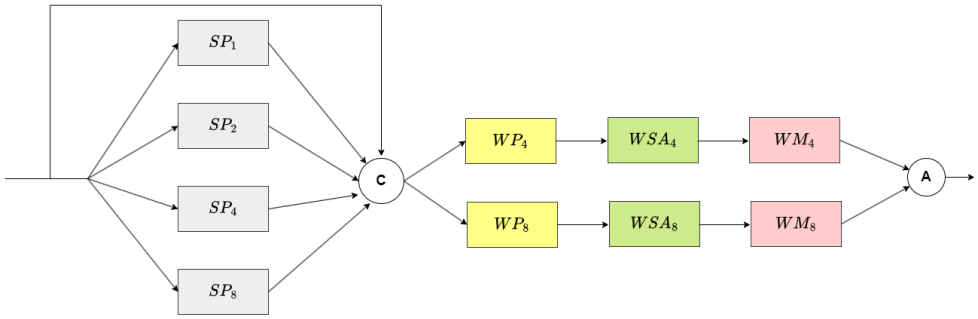


Figure 4.14: *The detail of Multi Context Local Self Attention Module. Here, SP_i is a spatial pooling operation that is calculated for different values of i. WP_8 and WP_4 are window partition operations with window-size of 8 and 4. WSA_8, and WSA_4 are window self-attention modules. WM_8 and WM_4 are window merge operations with window sizes 8 and 4.*

following equations.

$$R(x) = HA(x) + x \tag{4.18}$$

$$HA(x) = A_s(A_c(H(x)) * H(x)) * A_c(H(x)) * H(x) \tag{4.19}$$

Here, $R(x)$ is residual unit output when input x is passed, HA is the attentive nonlinear path, and H is a nonlinear path defined using a sequence of convolution, batch normalization, and Relu operation. We calculated A_s and A_c as proposed in [233].

4.2.1.3 Multi Context Local Self Attention module

We proposed a Multi Context Local Self-Attention Module (MCLSAM) as shown in Fig. 4.14. The computation of self-attention maps requires a lot of memory. We handle this using the generation of self-attention maps at small window inputs. We first generate multiple pooled features of input to get the multi-context information and concatenate it with the input feature. This combined input is partitioned into various small windows of sizes 4 and 8. Local self-attention maps are generated for both window size patches and merged individually. Finally, the combined self-attention map is calculated using the summation of merged maps for each size. We used the concept of window partitioning and local self-attention as proposed in [265]. For the feature map of size $B \times C \times H \times W$, the computational complexity calculations for multi-head self-attention (SA) and $MCLSAM$ are presented in

the following equations.

$$\begin{aligned} SA &= 4 \times H \times W \times C^2 + 2 \times (H \times W)^2 \times C \\ MCLSAM &= 8 \times H \times W \times C^2 + 2 \times \left(\frac{3M}{2}\right)^2 \times H \times W \times C \end{aligned} \tag{4.20}$$

The computational complexity of MCLSAM is linear proportional of $\mathcal{O}(H \times W)$ and SA has quadratic computational complexity of $\mathcal{O}(H^2 \times W^2)$. It shows that MCLSAM has the computational advantage over SA for input having a large $H \times W$ dimension. The qualitative comparison in GFLOPS is presented in Table 4.10.

$Window\_Size$	$GFLOPS$
$256 \times 256 + 256 \times 256$	107.01
$128 \times 128 + 128 \times 128$	29.70
$64 \times 64 + 64 \times 64$	10.38
$32 \times 32 + 32 \times 32$	5.54
$16 \times 16 + 16 \times 16$	4.34
$8 \times 8 + 8 \times 8$	4.03
$8 \times 8 + 4 \times 4$	4.01

Table 4.10: *Computational complexity for MCLSAM modules for various window sizes using input feature maps having a dimension of* $1 \times 48 \times 256 \times 256$.

Standard multi-head self-attention is calculated on original or downsampled resolutions ($\frac{1}{2}^{th}, \frac{1}{4}^{th}$) and it required significant GFLOPS. As shown in Table 4.10, MCLSAM used a window sizes of 8×8 and 4×4 which reduced significant FLOPS requirements compared to larger window size which is used in traditional self-attention maps calculations.

4.2.1.4 Attentive Reconstruction Module

We proposed a triad learning based attentive reconstruction module for the reconstruction of triads (input pairs and modulus of its difference) from low-resolution features from the backbone. The subpixels-based convolutions are used for single-image super-resolution in our task. We used the same backbone for this task with lower-resolution features of it. This module helped the network to learn the representation of the original image pair and its differences which is utilized for effective CD. Extracted features from the shared backbone are passed through ARM, consisting of two pairs of convolutions and spatial and

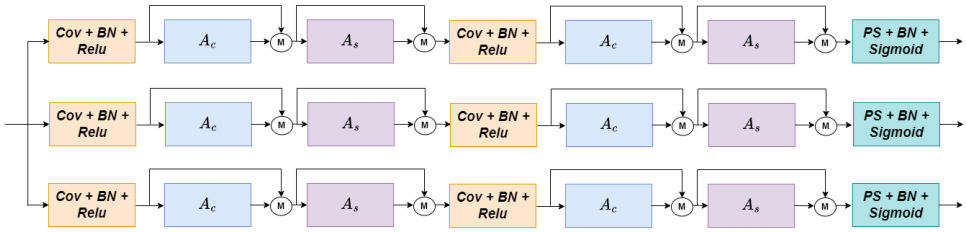


Figure 4.15: *The detail of Attentive Reconstruction Module. Here, M is the multiplication operation, and Conv+BN+Relu is the convolution followed by batch normalization and Relu operations. A_c and A_s are channels and spatial attention. PS+BN+Sigmoid is the pixel shuffle followed by batch normalization and sigmoid operations.*

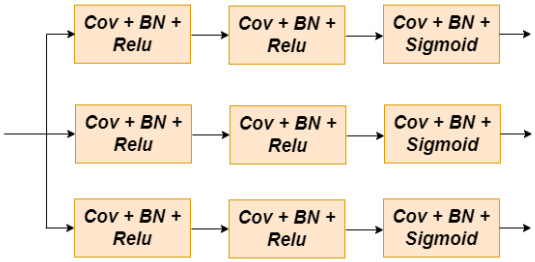


Figure 4.16: *The details of Auto Encoder Module. Here, Conv+BN+Relu is the convolution followed by batch normalization and Relu operations, and Conv+BN+Sigmoid is a sequence of convolution, batch normalization, and Sigmoid operations.*

channel attention units, followed by sub-pixel convolutions. We used three separate ARM for triad representations. We proposed this for super-resolution of features from $[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}]$ resolution streams of backbone. We reconstructed the low-resolution features using ARM, which reduces the noisy features in the reconstructed image because noisy features are also downsampled in low-resolution streams, but due to learning based reconstruction, it will be reduced in upsampling.

4.2.1.5 Auto Encoder Module

We proposed an Auto Encoder Module (AEM) as shown in Fig. 4.16 for the reconstruction of the triad from the original resolution features of the input of the backbone. We used a sequence of convolutions, batch normalization layers, and Relu activation functions. We used the sigmoid activation function before passing the final output from the module. AEM allows the network to learn feature reconstruction at actual resolution and enhance the feature representation capability of the network. We controlled the overall effect of AEM

in the network using weighted coefficients assigned to reconstruction loss for the AEM. We give a lesser value for AEM weight coefficients than ARM because the learning based reconstruction of low-resolution features allows better feature representation.

4.2.2 Composite Loss

Loss functions play an important role in the convergence of optimization problems. We proposed the use of a composite loss function that can be represented as follows:

$$Loss = L_{chg} + L_{sup} + L_{ae} \tag{4.21}$$

Where L_{chg}, L_{sup}, L_{ae} are losses for CD, super-resolution, and auto-encoder tasks. L_{chg} which is as used in [48,50], consist of cross entropy loss L_{ce} and dice loss L_{dc} as per following equation

$$L_{chg} = L_{ce} + L_{dc} \tag{4.22}$$

Loss for the super-resolution task is calculated for three resolutions for triad reconstruction. L_{sup} can be represented as

$$L_{sup} = L_{sup_1} + L_{sup_2} + L_{sup_3} \tag{4.23}$$

Here, $L_{sup_1}, L_{sup_2}, L_{sup_3}$ are super-resolution losses for three different resolutions output of the network. For resolution i, L_{sup_i} can defined as

$$L_{sup_i} = \alpha_i \times L_{diff} + \beta_i \times L_{lap_A} + \gamma_i \times L_{lap_B} \tag{4.24}$$

Here L_{diff} is the difference loss used for the reconstruction of pair difference and calculated using the mean square error loss function. L_A, L_B are reconstruction losses for input pair A and B, which are calculated using the Laplacian pyramid loss function, which was first used in [266]. We used weight value parameters $\alpha_i, \beta_i, \gamma_i$. For input x, target output t the Laplacian pyramid loss is calculated as

$$L_{lap} = \sum_{i=1}^{levels} |LapPyramid(x)_i - LapPyramid(t)_i| \tag{4.25}$$

We used three levels in our Laplacian pyramid loss function. For reconstruction of triad we utilized auto-encoder loss L_{ae} that is similar to L_{sup_i} and defined as

$$L_{ae} = \alpha_{ae} \times L_{diff} + \beta_{ae} \times L_{lap_A} + \gamma_{ae} \times L_{lap_B} \tag{4.26}$$

	α_i	β_i	γ_i
$i = \frac{1}{2}$	0.4	0.2	0.2
$i = \frac{1}{4}$	0.3	0.1	0.1
$i = \frac{1}{8}$	0.01	0.001	0.001
$i = ae$	0.2	0.05	0.05

Table 4.11: *Values of weighted parameters. Here, i is the resolution levels, and α, β, γ are the weighted parameters.*

4.2.3 Dataset

We used S2Looking [267], LEVIR-CD+ [53], SYSU [49], CDD [260], LEVIR-CD [53], WHU-BCDD [261], and DSIFN [256] datasets. The details of each dataset are presented in Table 4.12. If required, we crop the inputs to patches of 256 × 256. The CDD, SYSU, DSIFN, and S2Looking datasets contain multiple types of changes. The datasets of LEVIR-CD, LEVIR-CD+, and WHU-BCDD have changes of building only. For S2Looking datasets, ground truth is given for both directions between input pairs. We used only ground truth for changes between input first and input second.

4.2.4 Experiments

4.2.4.1 Training

We used hyperparameter values of 10 for batch size, 100 for the number of epochs, 0.0005 for weight decay, and 0.0001 for the initial learning rate. We used patches of size 256 × 256 during model training for each dataset due to the graphics memory limitations. We trained the model on the machine with an Nvidia P100 graphics card with 32 GB memory. The model is developed using the PyTorch library. We used benchmark splits in all datasets for

Dataset	Total image pairs	Train Set size	Validation Set size	Test Set size	Resolution (m)	Individual image size (pixels)	Change Instances
CDD [260]	16000	10000	3000	3000	0.03-1	256 × 256	-
LEVIR-CD [53]	637	445	64	128	0.5	1024 × 1024	31333
LEVIR-CD+ [53]	985	637		348	0.5	1024 × 1024	48455
WHU-BCDD [261]	2	1		1	0.07-0.3	15354 × 21243, 15354 × 11265	-
SYSU [49]	20000	12000	4000	4000	0.5	256 × 256	-
S2Looking [267]	5000	3500	500	1000	0.5-0.8	1024 × 1024	65920
DSIFN [256]	3940	3600	340	48	2	512 × 512	-

Table 4.12: *Details of various datasets used in our experiments. Here, - represents that number of change instances are not mentioned in the original work.*

performance comparison as utilized by other works. During the evaluation, we used the test time augmentations for calculations for final scores.

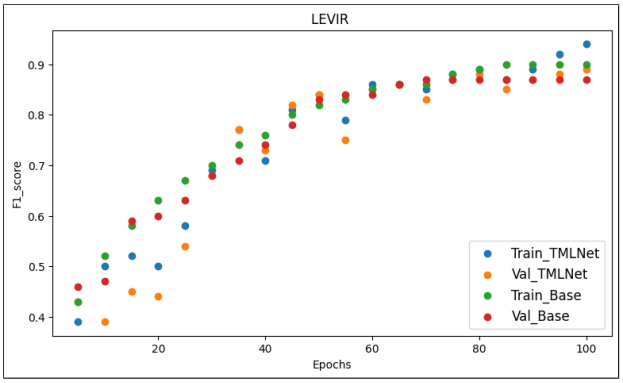
4.2.4.2 Evaluation

We use parameters of $F1\ Score$, $mIoU$, and overall accuracy (OA) for numerical comparison with other SOTA methods. The F1 Score (F_1) is calculated using precision (Pr) and recall (Rc) values, which are measured using true positive (tp), true negative (tn), false positive (fp), and false negative (fn).

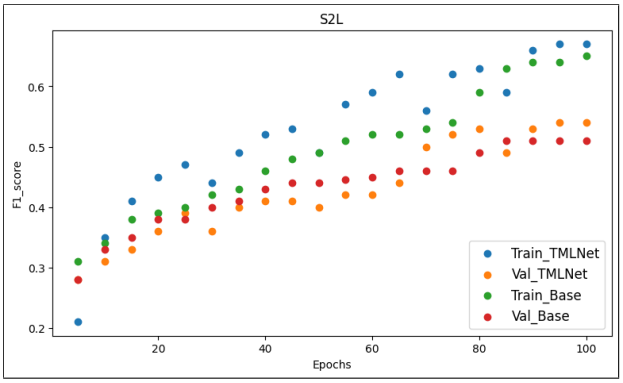
We used visual comparison for the qualitative evaluation of different model performances.

4.2.5 Result Analysis

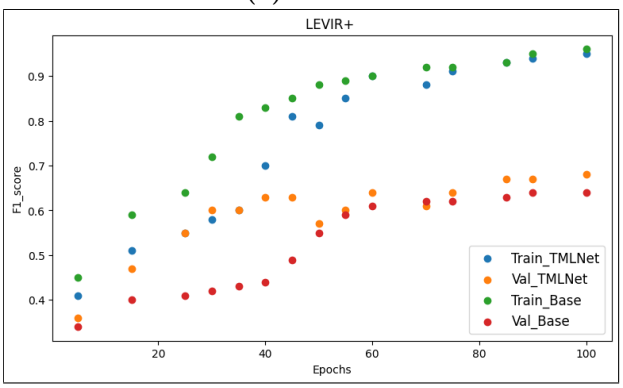
We performed quantitative and qualitative comparative analyses on seven benchmark datasets. For quantitative comparison, results of recently published SOTA methods are used. In qualitative comparison, we train five recent works of IFN [256], SNUNet [50], MFP [64], BIT [47], and DRMNet [48]. The visual results of these are compared with the proposed work.



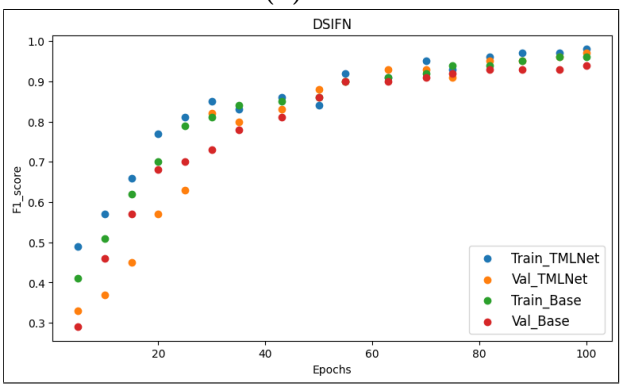
(a) *LEVIR*



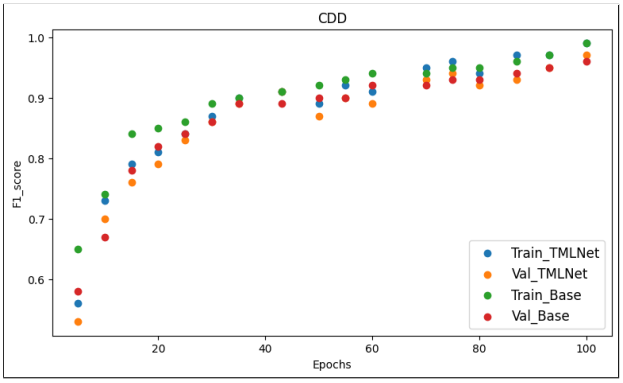
(b) *S2L*



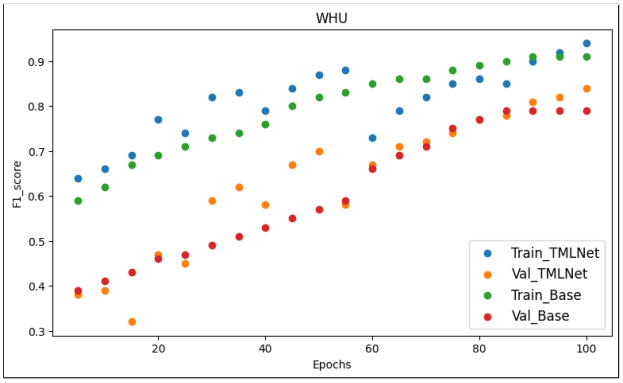
(c) *LEVIR+*



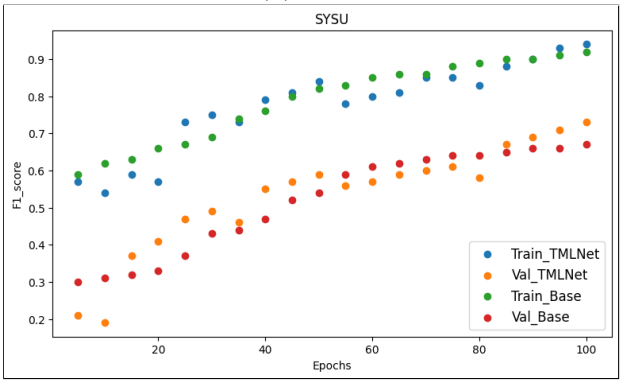
(d) *IFN*



(e) *CDD*



(f) *WHU*



(g) *SYSU*

Figure 4.17: *Overfitting analysis using validation loss curves. Here, Train_TMLNet and Val_TMLNet are training and validation F1 Score achieved by TMLNet. The base network F1 Score for training and validation is represented by Train_Base and Val_Base.*

	P_r	R_c	F_1	IoU	OA
IFN [256]	94.96	86.08	90.30		97.71
BA^{2}Net [255]	88.12	95.28	91.36		98.94
MFSNet [268]	95.70	93.30	91.58	81.55	
BSFNet [66]	90.5	93.3	91.9		98.10
FCL [269]			92.4		
DASNet [54]	93.2	92.2	92.7		98.2
ISNet [57]	95.18	94.43	94.80	90.12	98.78
JFSDNet [270]	98.75	92.60	95.55		99.07
SNUNet [50]	96.3	96.2	96.2		
MCCRNet [39]	97.52	95.32	96.41	96.07	99.25
AGCDetNet [56]	95.03	98.10	96.54	-	99.13
Bit [47] #	95.24	97.99	96.60	96.12	99.24
MFPNet [64] #	96.32	97.14	96.73	96.16	99.31
TSNet [63]	96.70	97.60	97.14	94.45	99.27
ADHR-CDNet [271]	99.50	95.00	97.20		
DRMNet [48]	97.92	98.49	98.20	96.46	99.57
TMLNet ours	98.11	98.69	98.40	96.85	99.62

Table 4.13: *Quantitative comparison results for CDD dataset. Here, the result with a # mark is produced by our implementation. The best values are displayed in red color.*

4.2.6 Quantitative Analysis

We achieved SOTA results for seven benchmark datasets for F_1, IoU, and OA. Triad learning with multi-context self-attention contributed to attaining the best results. The comparative analysis for each dataset is given below:

4.2.6.1 CDD

In Table 4.13, we presented the comparative analysis of our methods with other SOTA methods. For this dataset, the earlier best results are achieved by DRMNet [48] with the values of 97.92%, 98.49%, 98.20%, 96.46%, and 99.57% for P_r, R_c, F_1, IoU, and OA. The proposed TMLNet has achieved F_1, IoU, and OA of 98.40 %, 96.85%, and 99.62 % which is 0.20 %, 0.39 %, 0.05 % higher than the DRMNet [48]. We also achieved the highest Rc values with an improvement of 0.30%. The highest Pr value is attained by ADHRCDNet [271]. The improvement of 0.39% in the IoU value signifies the efficacy of TMLNet.

4.2.6.2 LEVIR-CD

For this dataset, the comparative result is shown in Table 4.14. The previous best results are produced by DRMNet [48] that has F_1, IoU, and OA values of 91.97%, 85.13%, and 99.19%. Our proposed work has achieved the values of P_r, R_c, F_1, IoU, and OA values of 93.69%, 90.54%, 92.09%, 85.34%, and 99.21%. It improved the F_1, IoU, and OA values by 0.12 %, 21%, and 0.03 % respectively from DRMNet [48]. We achieved the third-highest value of Pr and the fourth-highest value of Rc. The highest Pr and Rc values are achieved by ADHRCDNet [271] and TSNet [63]. The significant improvement in F_1, IoU values by TMLNet are due to MTL with effective far pixel relation mapping.

	P_r	R_c	F_1	IoU	OA
FCL [269]			83.10		
MSPSNet [272]	91.38	87.08	89.18		
IFN [256] #	90.65	88.01	89.31	80.24	98.98
BiT [47]	89.24	89.37	89.31	80.68	98.92
EGRCNN [46]	87.82	91.47	89.61		98.92
JFSDNet [270]	91.66	88.02	89.72		99.17
SNUNet [50]	90.61	89.01	89.80	81.49	98.97
ISNet [57]	92.46	88.27	90.32	82.35	99.04
MFPNet [64]			91.69		
MCCRNet [39]	89.91	89.62	90.71	91.13	99.24
AGCDetNet [56]	92.12	89.45	90.76	83.09	-
ChangeStar [273]	-	-	91.25	83.92	-
ADHR-CDNet [271]	94.60	88.50	91.40		
GPCM [274]	-	-	91.50		
ChangerEx [275]	92.97	90.61	91.77		
TSNet [63]	90.81	92.87	91.83	84.89	99.18
CEECNet [257]	93.81	89.92	91.83	84.89	-
DRMNet [48]	93.05	90.91	91.97	85.13	99.19
TMLNet Ours	93.69	90.54	92.09	85.34	99.21

Table 4.14: *Quantitative comparison results for LEVIR-CD dataset. Here, the result with a # mark is produced by our implementation. The best values are displayed in red color.*

4.2.6.3 SYSU

Table 4.15 presented the comparative results for this dataset. DRMNet [48] got the earlier best results for P_r, F_1, IoU, and OA with values of 84.55%, 80.53%, 67.39%, and 91.23%.

Our proposed work achieved P_r, R_c, F_1, IoU, and OA values of 82.69%, 79.25%, 80.93%, 67.97% and 91.19%. We achieved the improvement of 0.3 %, and 0.58 % % in F_1, and IoU values. We achieved the second-best Pr and OA values and the fifth-best Rc value. The highest Pr and OA are achieved by DRMNet [48], and the highest Rc values are achieved by MSPSNet [272]. TMLNet attained notable improvements in F_1 and IoU value due to the joint approach of triad learning with multi-context self-attention.

	P_r	R_c	F_1	IoU	OA
IFN [256] #	76.11	78.45	77.26	64.81	89.21
MSPSNet [272]	76.14	79.93	77.39		
ISNet [57]	80.27	76.41	78.29	64.44	78.29
MFPNet [64] #	77.85	79.15	78.50	65.07	89.83
Bit [47] #	77.97	79.76	78.85	65.11	90.03
SNUNet [50]	78.16	79.68	78.92	65.18	89.96
DRMNet [48]	84.55	76.86	80.53	67.39	91.23
TMLNet Ours	82.69	79.25	80.93	67.97	91.19

Table 4.15: *Quantitative comparison results for SYSU dataset. Here, the result with a # mark is produced by our implementation. The best values are displayed in red color.*

4.2.6.4 LEVIR-CD+

The results comparison is presented in Table 4.16 for this dataset. BIT [47] has achieved the previous best value of 82.80% for F_1. The proposed work achieved the F_1, IoU, and OA values of 83.06 %, 71.04 %, 98.64 % and improved the F_1 by 0.26 % from the previous work [47]. TMLNet achieved the second highest Pr value preceded by CDNet [58]. It got the fourth highest Rc value which is preceded by STANet [53], BIT [47], SNUNet [50].

4.2.6.5 S2Looking

The comparative analysis is presented in Table 4.17. Bit [47] got the previous best values of 61.85% for F_1. The proposed work achieved the values of 63.89%, 46.94% and 99.19% for F_1, IoU, and OA respectively. It got an improvement of 2.04 % in F_1 value from the previous work [47]. TMLNet also gets the second highest value of Rc preceded by DRMNet [48] and the third highest value of Pr. TMLNet achieved significant improvement in F_1 value,

	P_r	R_c	F_1	IoU	OA
FC-EF [251]	61.30	72.61	66.48		
IFN [256] #	64.16	70.34	67.11	62.89	97.67
MFPNet [64] #	65.36	81.64	72.60	63.21	97.44
DTCDSCN [51]	80.36	75.03	77.60		
STANet [53]	74.62	84.54	79.31		
DRMNet [48] #	80.90	77.99	79.42	65.86	98.35
CDNet [58]	88.96	73.45	80.46		
SNUNet [50]	78.69	82.31	80.45	67.31	98.37
Bit [47]	82.74	82.85	82.80	-	-
TMLNet Ours	84.10	82.05	83.06	71.04	98.64

Table 4.16: *Quantitative comparison results for LEVIR-CD+ dataset. Here, the result with a # mark is produced by our implementations. The best values are displayed in red color.*

showcasing the proposed approach's effectiveness.

	P_r	R_c	F_1	IoU	OA
FC-EF [251]	81.36	08.95	07.65		
SNUNet [50]	33.61	35.29	34.43	20.79	98.37
STANet [53]	38.75	56.49	45.97		
IFN [256] #	67.54	48.91	56.73	35.78	98.78
DTCDSCN [51]	68.58	49.16	57.27		
MFPNet [64] #	67.89	52.63	59.29	39.12	98.93
CDNet [58]	67.48	54.93	60.56		
DRMNet [48] #	62.44	60.92	61.67	44.58	99.08
Bit [47]	72.64	53.85	61.85	-	-
TMLNet Ours	68.65	59.74	63.89	46.94	99.19

Table 4.17: *Quantitative comparison results for S2Looking dataset. Here, the result with a # mark is produced by our implementations. The best values are displayed in red color.*

4.2.6.6 WHU-BCDD

The comparison is presented in Table 4.18 for this dataset. Due to the use of different test split settings, we train IFN [256], SNUNet [50], MFP [64], BIT [47], and DRMNet [48] on this dataset. The results produced by these are used for quantitative comparison. The previous best values of 87.93%, 86.11%, 75,61%, and 99.01% for P_r, F_1, IoU, and OA is achieved by DRMNet [48]. The proposed work achieved the F_1, IoU, and OA value s of 87.79 %, 78.23 %, and 99.12 % which improved the results by 1.68 %, 2.62 %, and .11 % from [48]. Our work has achieved the best results for P_r, F_1, IoU, and OA and the second

best value for R_c preceded by BIT [47].

	P_r	R_c	F_1	IoU	OA
MFPNet [64] #	36.25	81.25	50.13	33.44	94.13
IFN [256] #	80.06	84.95	82.43	70.12	98.68
SNUNet [50]	85.25	81.09	83.12	71.12	98.80
Bit [47] #	81.71	90.59	85.92	75.30	98.91
DRMNet [48]	87.93	84.37	86.11	75.61	99.01
TMLNet Ours	89.27	86.34	87.79	78.23	99.12

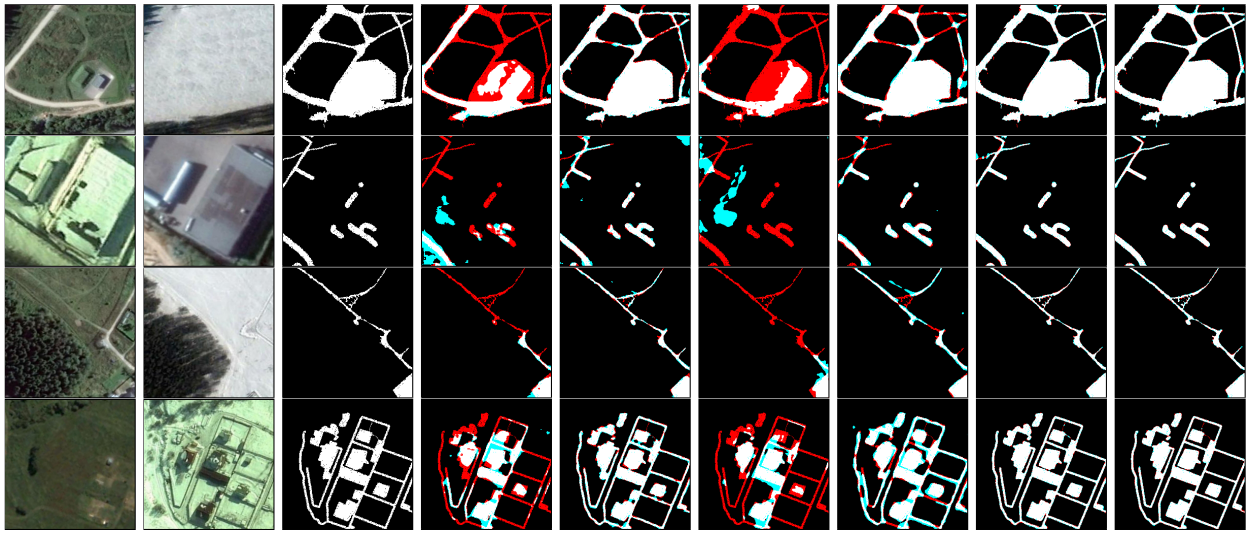
Table 4.18: *Quantitative comparison results for WHU BCDD dataset. Here, the result with a # mark is produced by our implementations. The best values are displayed in red color.*

4.2.6.7 DSIFN

Table 4.19 shows this dataset's results. MFPNet [64] has the previous best results for F_1 with the value of 69.14%. We are able to achieve F_1 values of 68.68 %, which is 0.69 % better than earlier work [64]. We trained DRMNet [48] on this dataset for performance comparison. The values of F_1 and IoU of the proposed model are 1.49% and 2.83% higher than the DRMNet [48]. We improved the OA values by 1.34% with a value of 89.87%. Our model has achieved the highest P_r value and the fourth highest value of R_c, which is lesser than MapsNet [276] and IFN [256]. TMLNet attained significant improvement in F_1 value that showcases the efficacy of the proposed approach.

	P_r	R_c	F_1	IoU	OA
Unet++ + MSOF [252]	59.83	65.91	62.73	-	86.68
SNUNet [50] #	65.11	69.33	67.15	48.87	87.11
IFN [256]	67.11	67.54	67.33	-	88.40
MapsNet [276]	64.51	70.48	67.36	-	88.86
Bit [47] #	69.97	65.01	67.39	49.73	88.51
DRMNet [48] #	70.39	65.11	67.65	50.01	88.53
MFPNet [64]	-	-	68.45		
TMLNet ours	71.64	66.81	69.14	52.84	89.87

Table 4.19: *Quantitative comparison results for DSFIN dataset. Here, the result with a # mark is produced by our implementations. The best values are displayed in red color.*



(a) *I1* **(b)** *I2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.18: *Qualitative comparison for the patches of the CDD test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*

4.2.7 Overfitting Analysis

In figure 4.17, a detailed analysis of avoidance of overfitting using multitask is presented. For each dataset, TMLNet and Base network training and validation step F_1 are compared for 100 epochs. Multitask network graph avoided the overfitting. It is reflected through the sudden decrease in F_1 for some epochs, which is followed by increasing trends of F_1. This could be because multitask network adjusted the learning and resulted into better performance on validation and test sets. Base network without MTL converges faster to higher F_1 values, but the highest validation F_1 is saturated after some time. These trends follow in all subfigures of each dataset training validation F_1 curve.

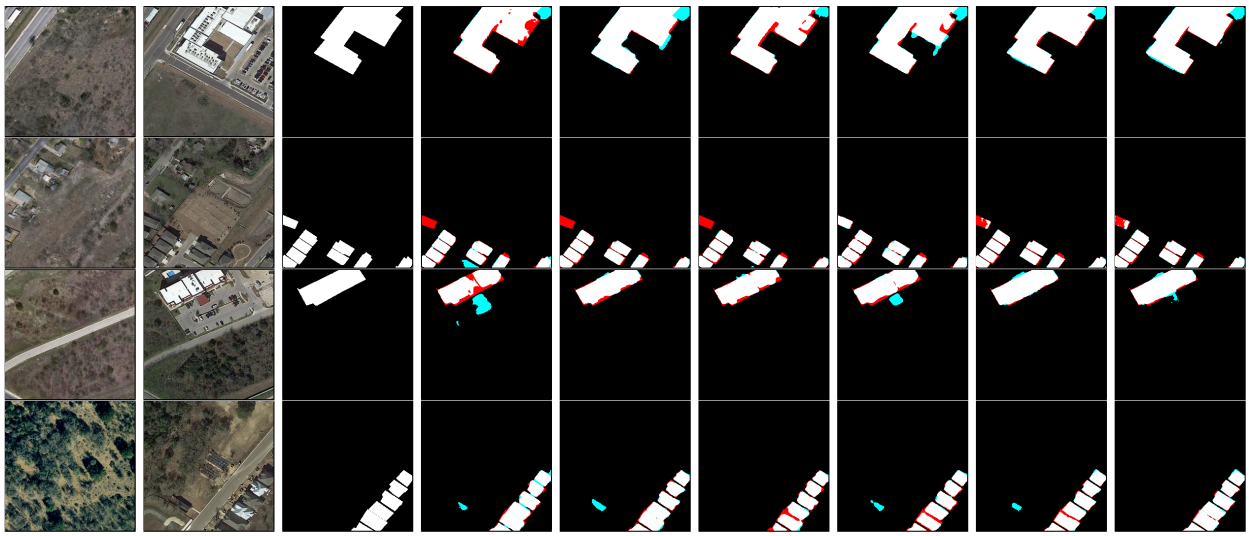
4.2.8 Qualitative Analysis

For qualitative analysis, we compared the visual result produced by the proposed work with recent best published works like DRMNet [48], BIT [47], IFN [256], SNUNet [50], MFP [64]. We trained each network for 150 epochs, with training being stopped when the model was unable to improve the result further. In Fig. 4.18, CDD dataset visual results

are compared. The TMLNet generated the closest results from the ground truth. The most comparative results from TMLNet are generated by DRMNet [48], and SNUNet [50] for this dataset. For the LEVIR dataset, the visual comparison is shown in Fig. 4.19. It indicates that the TMLNet gives giving most accurate change map compared to others with the least detections of false change. Other methods also show comparative results for this dataset. In Fig. 4.20, visual analysis for the LEVIR-CD+ dataset is presented. The TMLNet and BIT [47] produced the most accurate results, while DRMNet [48] missed the actual changes for some patches, and false changes are detected by SNUNet [50] and IFN [256]. Fig. 4.21 compared the visual results for the S2Looking dataset. Each model faced difficulty in precisely generating a change map for this dataset. TMLNet, BIT [47] produced the most accurate results. IFN [256] and SNUNet [50] missed a few actual changes and identified some spurious changes. The SYSU dataset results are shown in Fig. 4.22. TMLNet, DRMNet [48], and SNUNet [50] produced the most accurate results. MFP [64] and BIT [47] missed actual changes more than others. Fig. 4.23 presented the visual results for WHU-BCDD datasets. TMLNet, BIT [47], and SNUNet [50] produced comparatively better results. All models detected some spurious changes and missed out few actual changes. The visual results for the DSIFN dataset are shown in Fig. 4.24. All models performed poorly for a few cases in this dataset. TMLNet is able to avoid spurious changes, which other models failed. It also missed several actual changes. Other models overestimated the changes and detected spurious changes in most of the cases. In the proposed scheme, intuitively, the use of multi-scale attention helped TMLNet in better feature representations which are further enhanced by mapping of multi-context far pixel relations and MTL with laplacian loss that allowed the retention of finer details in reconstructed images and avoided overfitting.

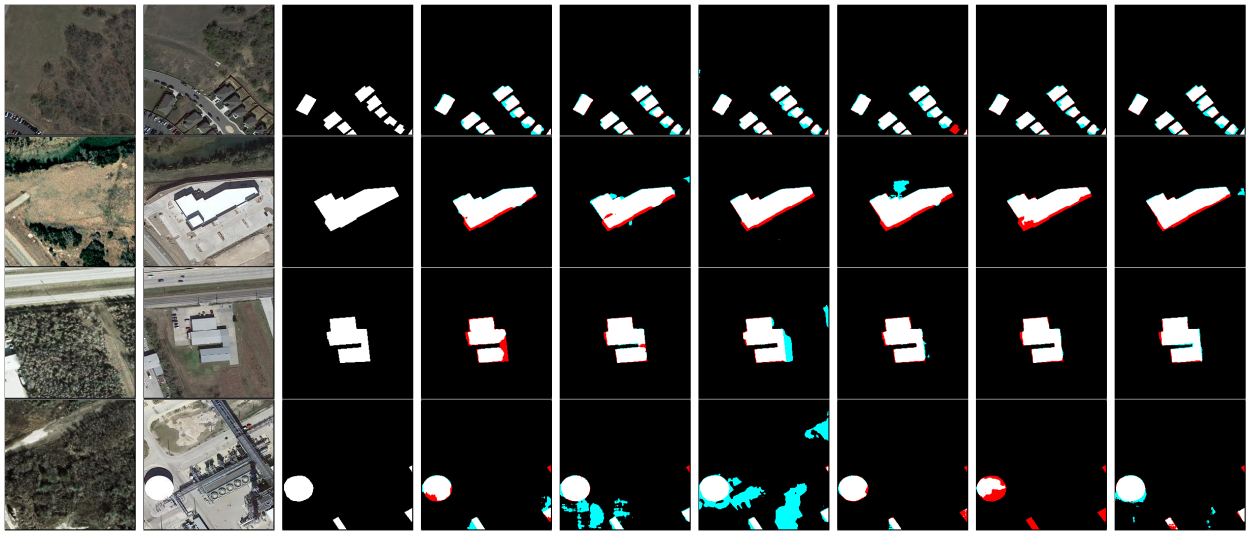
4.2.9 Ablation Study

We used the LEVIR-CD dataset in all our ablation studies. We train each network for 100 epochs in each configuration setting. We did ablation studies for the backbone module, loss functions, and network components. We used a multiscale backbone with four input



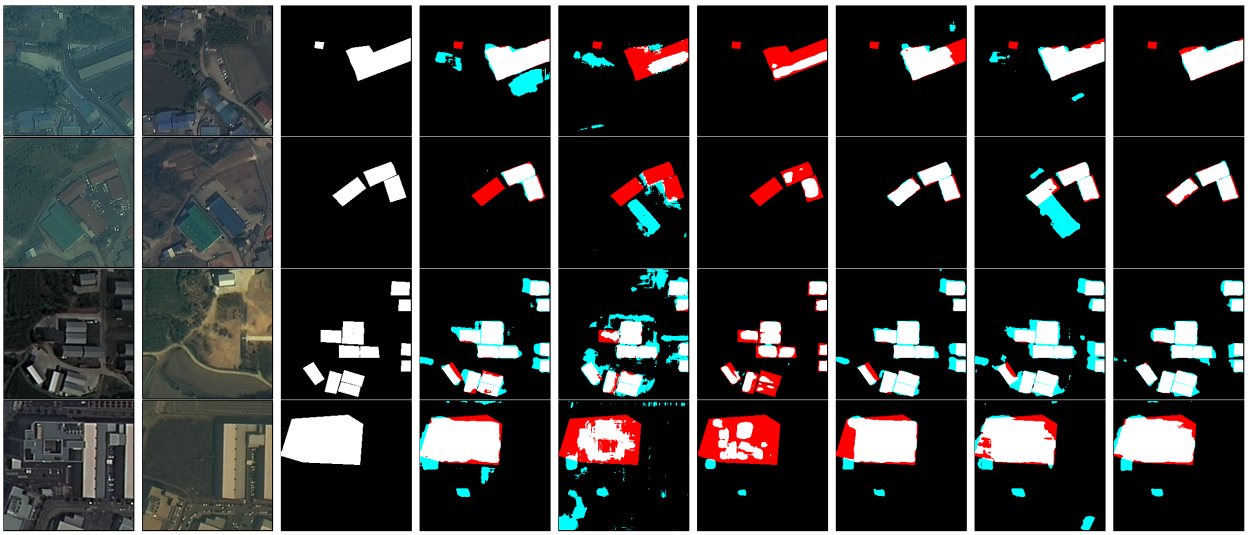
(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.19: *Qualitative comparison for the patches of the LEVIR test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*



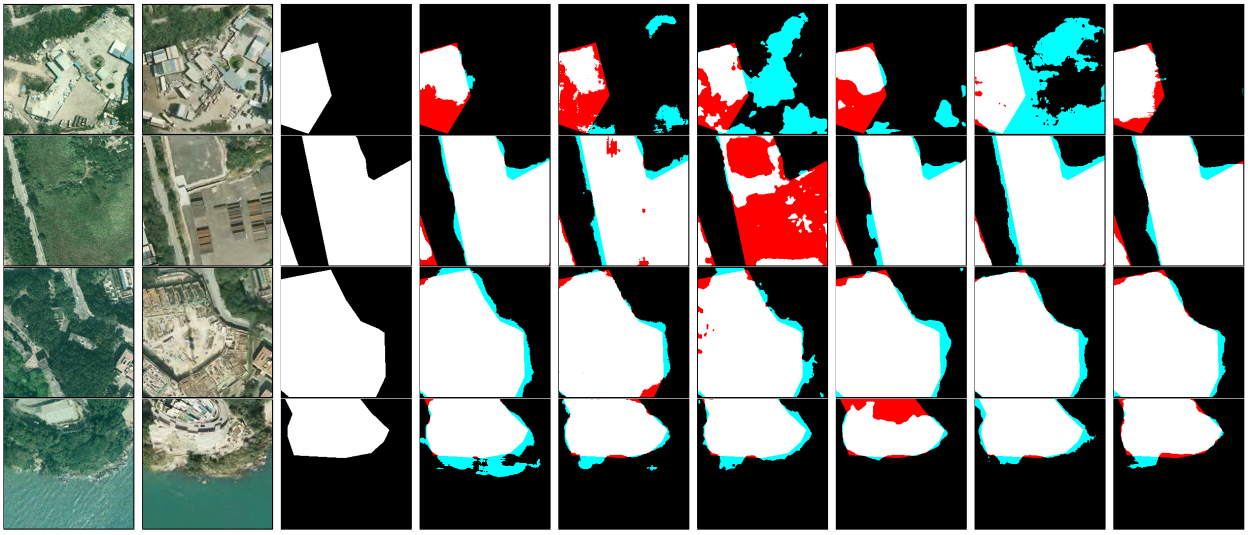
(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.20: *Qualitative comparison for the patches of the LEVIR-CD+ test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*



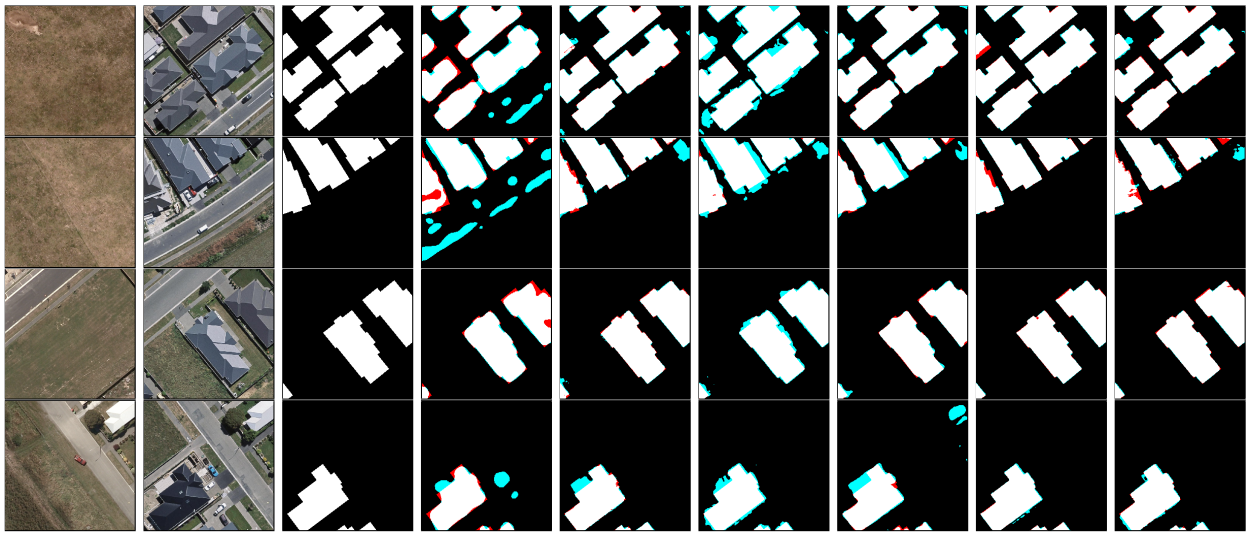
(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.21: *Qualitative comparison for the patches of the S2Looking test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*



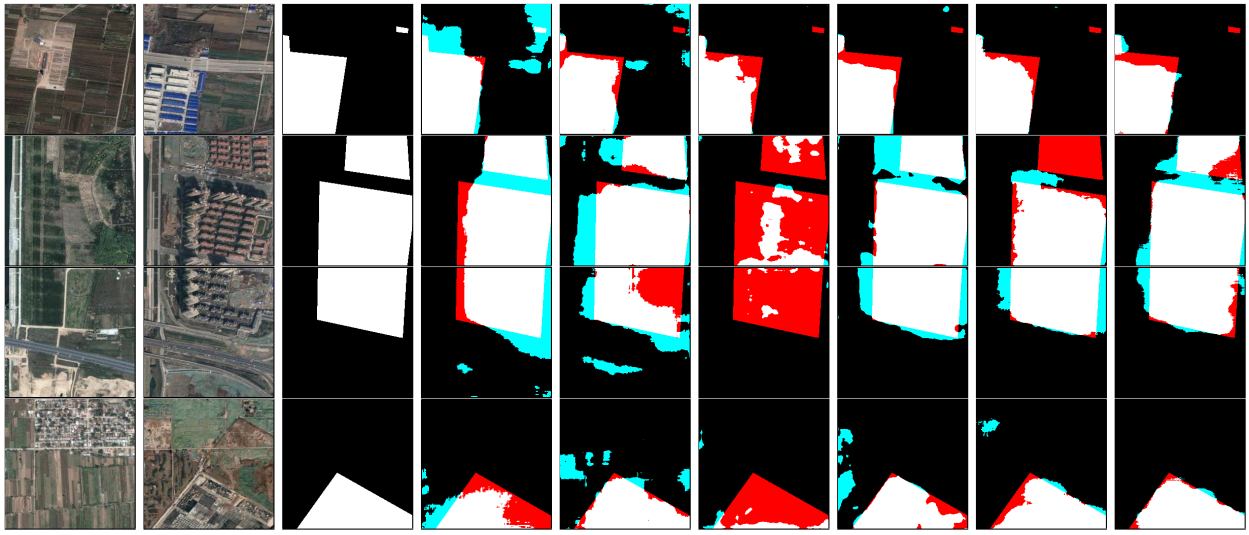
(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.22: *Qualitative comparison for the patches of the SYSU test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*



(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.23: *Qualitative comparison for the patches of the WHU test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*



(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *MFP* **(e)** *SNU* **(f)** *IFN* **(g)** *Bit* **(h)** *DRM* **(i)** *TML*

Figure 4.24: *Qualitative comparison for the patches of the DSIFN test dataset. Here, the size of each patch is* 256×256 *pixels. The white color represents the actual change, the red color represents the change missed by models, and the aqua color represents false changes detected by models.*

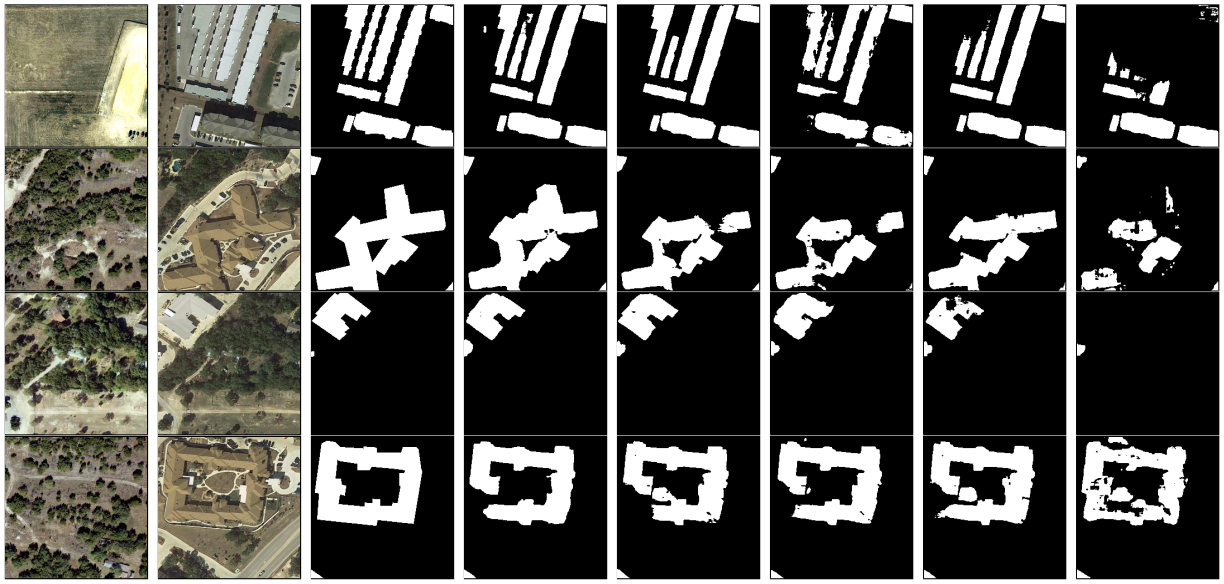
resolution levels.

4.2.9.1 Ablation Study for backbone

We presented an ablation study for the backbone in Table 5.3. We started with a modified HRNet [45] backbone with only full high-resolution blocks. We will refer to it as a modified block (MB). We use two of these blocks in our backbone. We started with 48 initial numbers of channels in the backbone and the Early+difference fusion strategy for input pairs as used in DRMNet [48]. After training it gave F_1, IoU, and OA of 89.00 %, 80.19 %, and 98.89 % respectively. We further enhanced this modified block by using dense connections that resulted in the improvement of 0.21 %, 0.33 %, and 0.02 % in F_1, IoU, and OA. Later we incorporated topdown connections in the densely connected modified block that further improved the F_1, IoU, and OA with the values of 89.39 %, 80.81 %, and 98.92 %, respectively. Lastly, we replaced each residual unit of the modified block with an attentive residual block that contains channel-wise and spatial attention within the residual unit. We achieved the respective values of 89.55 %, 81.07 %, and 98.94 %. It is the final backbone module we used in this work and refers to as the Enhanced Backbone Module (EBM).

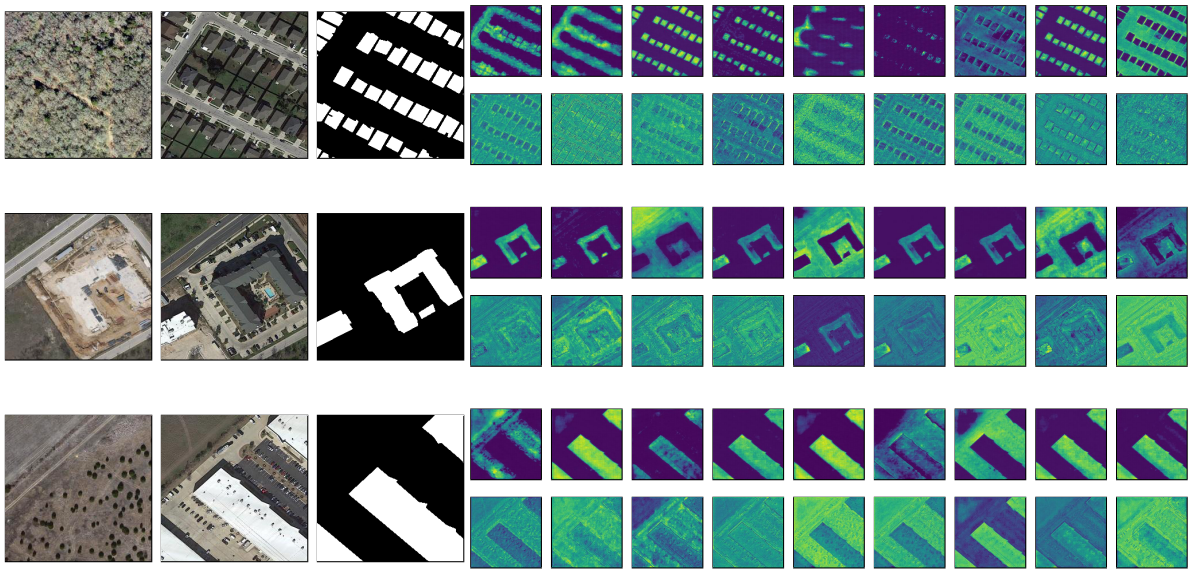
4.2.9.2 Ablation Study for loss function

We experimented with different loss functions as presented in Table 4.22. We trained TML-Net with $Cross-entropy$ and $MeanSquareError$ loss functions for change map and reconstruction of features, respectively, which resulted in the values of 91.76%, 84.72% for $F1Score$ and IoU. Further, we include the $Dice$ loss, which improves the values by 0.21% and 0.13%, respectively. Finally, we incorporated the $Laplacian$ loss function for feature reconstruction and retrained the model achieved the final values of 92.09% and 85.34% for $F1Score$ and IoU.



(a) *Input 1* **(b)** *Input 2* **(c)** *Label* **(d)** *C4* **(e)** *C3* **(f)** *C2* **(g)** *C1* **(h)** *C0*

Figure 4.25: *C0 = MB; C1 = C0 + MC; C2 = C1 + T; C3 = C2 + TA; C4 = C3 + A; Qualitative comparison for ablation study on patches of LEVIR dataset. Here, the size of each patch is* 256×256 *pixels. MB is the enhanced backbone module, MC is the modified local self-attention module, T is triad reconstruction on* $\frac{1}{2}$ *resolution stream, TA is triad on all low-resolution streams, and A is auto encoder module.*



(a) *I1* **(b)** *I2* **(c)** *L* **(d)** *f1* **(e)** *f2* **(f)** *f3* **(g)** *f4* **(h)** *f5* **(i)** *f6* **(j)** *f7* **(k)** *f8* **(l)** *f9*

Figure 4.26: *Feature map visualization for LEVIR-CD dataset. Here, I1,I2 and L are input pairs and ground truth. Randomly selected feature maps are shown in* $f1$ *to* $f9$*. The top sub row shows the feature maps for the output features of MCLSAM and the bottom sub row represents input feature maps of MCLSAM.*

4.2.9.3 Ablation Study for network component

The detailed study on the importance of each module of the TMLNet is presented in Table 4.21. From the previous section, we found that EBM is giving better $F1Score$ and IoU as represented in Table 5.3 compared to other backbone blocks. We added MCLSA at the end of EBM, which improved by .14% and 0.53% in $F1Score$ and IoU. This change improved the IoU score significantly. Further, the network is improved with the inclusion of triad reconstruction with ARM for $\frac{1}{2}$ resolution stream. This achieved the $F1Score$ and OA of 90.83 % and 83.19 % respectively. This change adds the improvement of .69% and 1.14% in $F1Score$ and OA. Motivated by this result, we extended TRM for all low-resolution streams and achieved 91.82% and 84.91% for $F1Score$ and IoU with the improvement of 0.99% and 0.72%. This ablation study's second biggest improvement was in the IoU values. Finally, we added AEM for the original resolution feature map reconstruction, which produced the values of 92.09% and 85.34% for $F1Score$ and IoU. It increases the performance matrices by the values of 0.27%. and 0.43%. The visual analysis of the ablation study is presented in Figure 4.25. This figure shows that the inclusion of each component of the proposed network has contributed to the enhancement of the change map results. Feature map visualization for 9 randomly selected features (out of 48 features) is presented in Figure 4.26 for input and output features of MCLSAM. The output features of MCLSAM are more correlated to the change maps which are not prominent in input features of MCLSAM which showcased the effectiveness of MCLSAM.

4.2.9.4 Computational complexity comparison

In Table 4.23, we compared the computational resource requirement of the number of parameters and floating point operations by DRMNet [48], SNUNet [50], MFPNet [64], IFN [256], and our network. The SNUNet required the least number of parameters. Our proposed network required 33.6 M trainable parameters, which is lesser than DRMNet [48], MFP-Net [64], and IFN [256]. For GFlops, operations required by the proposed network have the third least requirement.

	P_r	R_c	F_1	IoU	OA
MB	90.34	87.71	89.00	80.19	98.89
MB + D	90.21	88.24	89.21	80.52	98.91
MB + D + TD	90.14	88.65	89.39	80.81	98.92
AB + D + TD	90.32	88.79	89.55	81.07	98.94

Table 4.20: *Ablation Study for the backbone. Here, MB is the modified-backbone, D is the dense connection, TD is the topdown connection, and AB is the modified-backbone with attentive residual connections.*

	P_r	R_c	F_1	IoU	OA
EBM	90.32	89.67	90.00	81.52	99.00
EBM + MCLSA	91.16	89.15	90.14	82.05	99.00
EBM + MCLSA + TR-1	92.48	89.23	90.83	83.19	99.08
EBM + MCLSA + TR-All	92.81	90.84	91.82	84.91	99.13
EBM + MCLSA + TR-All + AEM	93.69	90.54	92.09	85.34	99.21

Table 4.21: *Ablation Study for different network components. Here, EBM is the enhanced backbone module, MCLSA is the modified local self-attention module, TR-1 is triad reconstruction on $\frac{1}{2}$ resolution stream, TR-All is triad on all low-resolution streams, and AEM is auto encoder module.*

4.3 Summary

In this contributory chapter we presented details of two models (DRMNet, TMLNet). In DRMNet, a multi-tasking deep learning model comprised of a backbone network, a multi-scale attention module, and a sub-pixel convolution based deconvolution module. The proposed model can predict change map and image difference in parallel and uses two loss functions, hybrid loss and MSE loss. Initial feature fusion strategy with modules of features difference is applied in our network, and it has outperformed the recent best published works. A detailed study is presented for justifying the proposed loss functions. An ablation study is also presented to highlight the contributions of the different modules of the proposed architecture. A comprehensive set of experiments reveal that the proposed model has achieved the SOTA results for CDD, SYSU, and LEVIR-CD datasets. We also have set benchmark results for the BCDD dataset for future comparison.

In second work, we proposed a MTL network, TMLNet, that used triad reconstruction for enhanced change map generation. We have used the multiobjective based learning with a shared backbone which allowed a better depiction of features. The use of an attention

	F_1	IoU
MSE + CE	91.76	84.72
MSE + CE + DC	91.97	84.85
MSE + LP+ CE + DC	92.09	85.34

Table 4.22: *Loss functions Comparison. MSE is Mean Square Error, CE is cross-entropy loss, DC is Dice loss, and LP is Laplacian loss.*

	Parameters	GFlops
IFN [256]	50.44M	82.3
MFPNet [64]	85.9M	128.8
DRMNet [48]	35M	189.1
TMLNet	33.6M	123.2
SNUNet [50]	12.3M	54.8

Table 4.23: *Computational Comparison. Here, the number of parameters are counted in millions, and the number of floating point instructions is measured in GigaFlops.*

mechanism further improves this enhanced feature representation. The spatial and channel attention focused on essential features, which were further enhanced by local self-attention. The utilization of the Laplacian pyramid loss for feature reconstruction helped in the retention of low-level structures in the reconstructed image. We conducted extensive experiments, and the results verify that our proposed work has achieved SOTA results in seven benchmark datasets. Detailed ablation studies on various elements of the proposed network and other parameters are presented. We will share the implementation code and trained weights of TMLNet for comparative evaluation and future work.

The next contributory chapter presents the details of proposed models for the Image Translation (IT) task. Two important IT tasks of Multi modal (MM) SAR to RGB image translation and multi-temporal image Super Resolution (SR) are chosen as these are required for all weather mapping and better spatial interpretation.

"What you seek is seeking you."

∼ Rumi

5

Multimodal and multitemporal models for image translation

In this chapter, we discuss the third contribution to the Image Translation (IT) task. IT as a Remote Sensing Vision Task (RSVT) converts earth observation data from one type to another. This conversion can be between different modalities (Multi modal (MM) translation), different spatial resolutions (super-resolution), and different spectral resolutions (spectral synthesis). Synthetic Aperture Radar (SAR) to optical conversion enables all weather mapping because optical images are unavailable during cloudy conditions using MM IT. Multi-frame Super Resolution (SR) generates high spatial resolution images from a sequence of low-resolution inputs using multitemporal IT. We proposed two Deep Learning

(DL) models named Encoder Decoder based Conditional GAN (EDCGAN)) and Multi Context Dense Network (MCDNet) for MM and multitemporal IT. The significant contributions of this chapter are summarized as follows:

- An encoder-decoder-based model, named EDCGAN, is proposed for IT from SAR to RGB with a multi-scale attentive discriminator for IT.
- The proposed MCDNet model utilizes multiple spatial-space, self-attention, and dense residual attention for multi-frame SR.

The rest of the chapter is organized as follows. In Section 5.1 the details of the proposed EDCGAN architecture, experiments, and results are shown. Section 5.2 described the proposed model named MCDNet and experimentation and results details. The chapter is concluded in Section 5.3.

5.1 EDCGAN

In cloudy conditions, an optical sensor can not capture the ground reality, and it can be captured using a SAR sensor. The visualization of SAR images is difficult due to the presence of various noises. SAR to optical IT is a task that focuses on converting SAR images to optical images for better visualization. The proposed architecture is an encoder-decoder-based conditional Generative Adversarial Network (GAN) with multi-scale attentive discriminator to generate precise SAR to RGB IT. In addition, we have used residual connections, and spatial & channel-wise attention for better feature representation. The major contributions are as follows:

1. An encoder-decoder based multi-scale attentive discriminator with auxiliary loss is proposed to enhance the discriminative ability.
2. An encoder-decoder based generator is devised with multi-scale attention and residual connections to generate finer translated images.

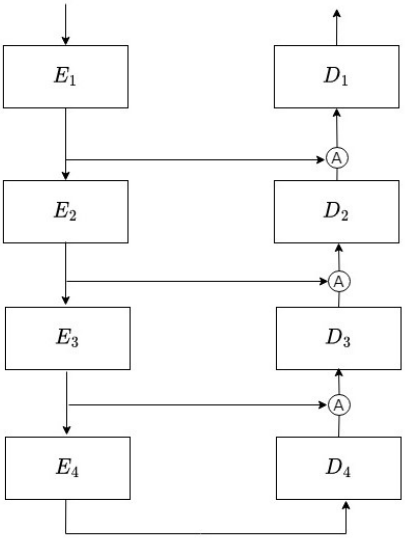
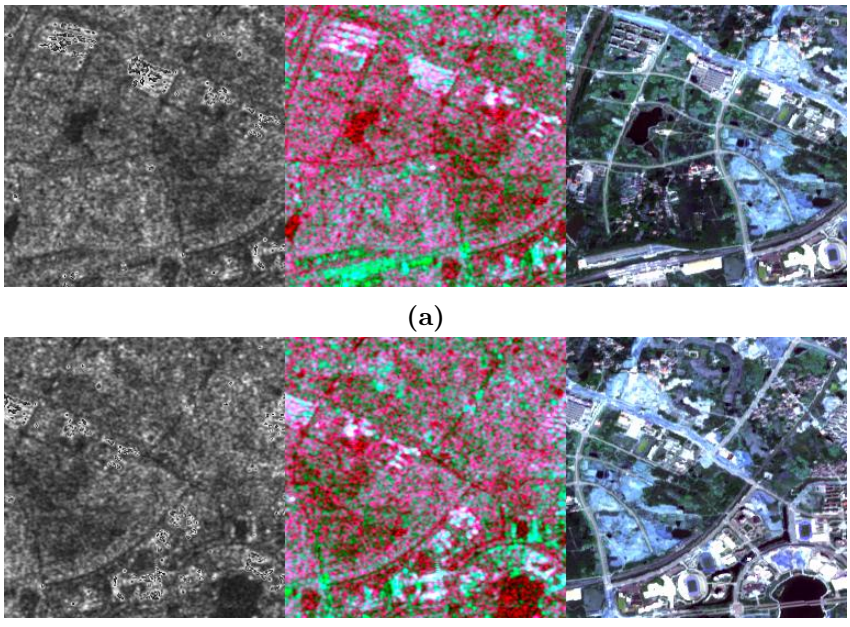


Figure 5.1: *The architecture of the encoder-decoder network used in our generator and discriminator network. Here, E is the encoder block, and D is the decoder block.*



(a)

(b)

Figure 5.2: *Visualization of patches of raw SAR image (left), pre-processed SAR image (middle), and Sentinel 2 produced RGB image (right) of size* 256×256 *are shown.*

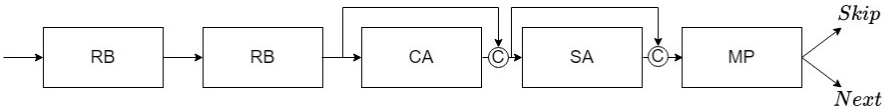


Figure 5.3: *Block diagram of the encoder. Here, RB is the residual block, CA is channel attention, SA is spatial attention, and MP is the max-pooling operation. Skip is the skip connection to the decoder block, and Next is a connection to the next encoder block.*

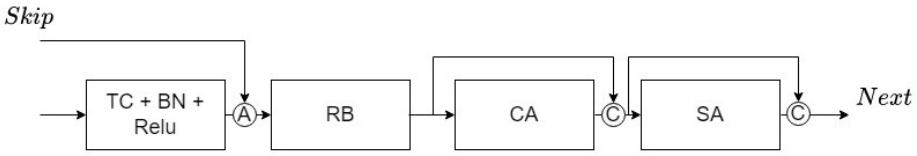


Figure 5.4: *Block diagram of the decoder.*

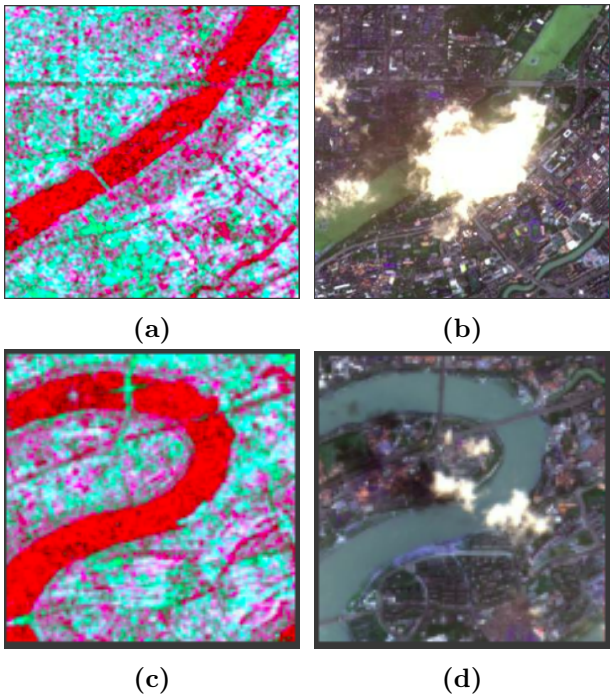


Figure 5.5: *The visualization of Sentinel 2 images in cloud condition.*

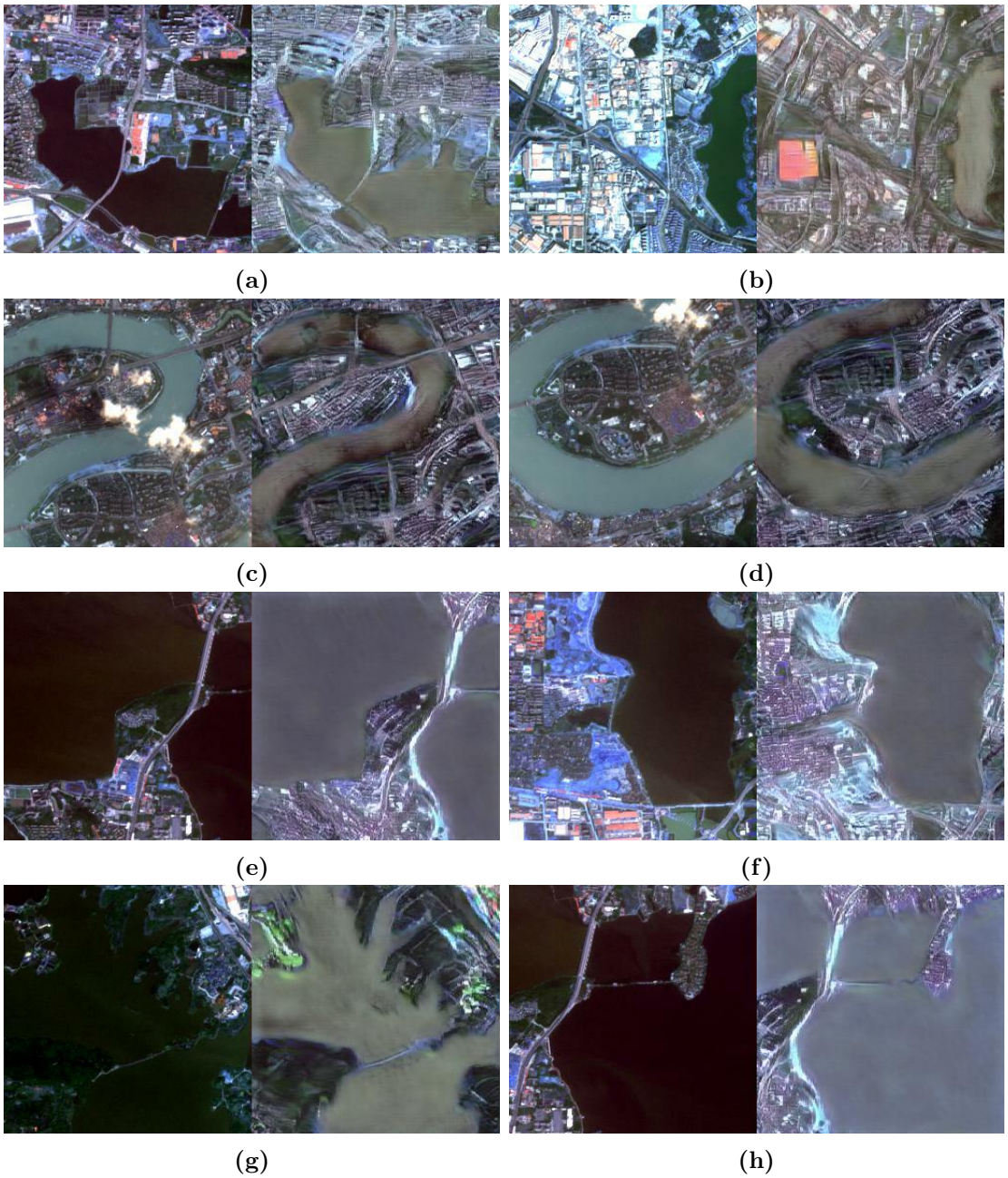


Figure 5.6: *Comparision of the actual sentinel-2 image with the generated images. The left portion contains Sentinel-2 data, and the right portion is predicted image.*

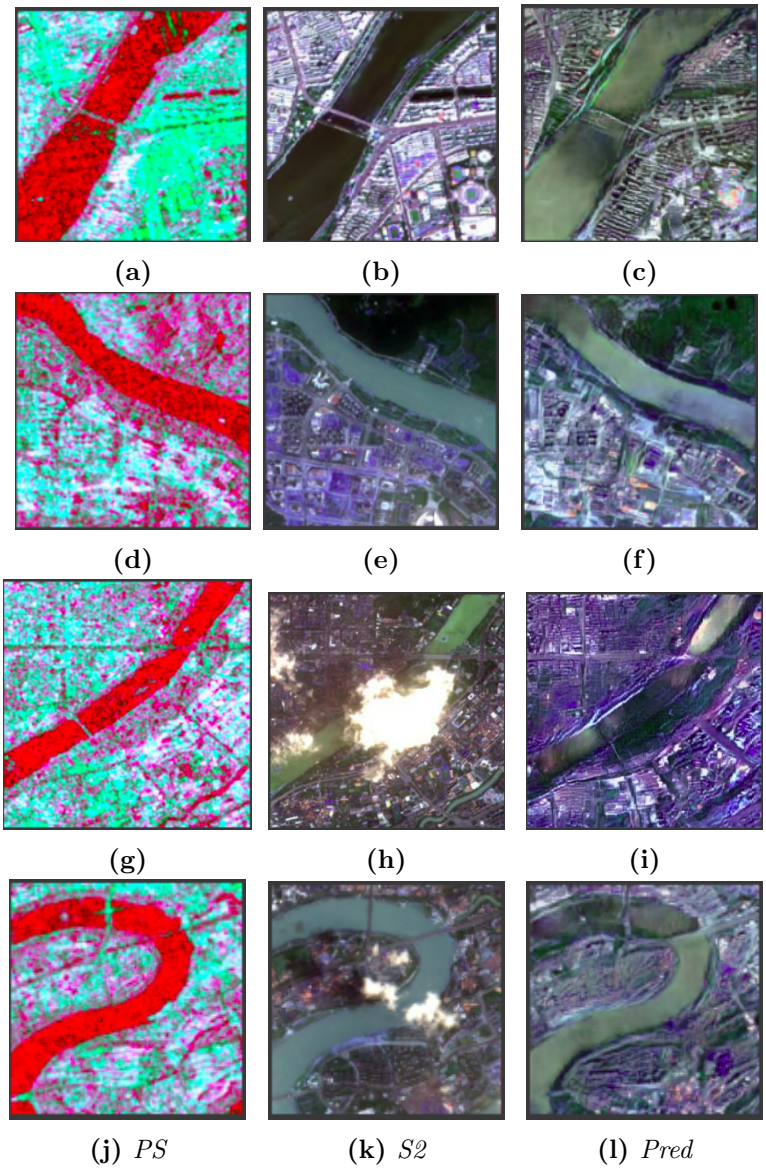


Figure 5.7: *Qualitative analysis of the SAR to RGB translation performed by the proposed model. Here PS is the processed SAR input image, S2 is the corresponding sentinel-2 image, and Pred is the predicted output image. Bridges and other structures are preserved in 5.7c, 5.7f. Clouds have been removed in 5.7i, 5.7l*

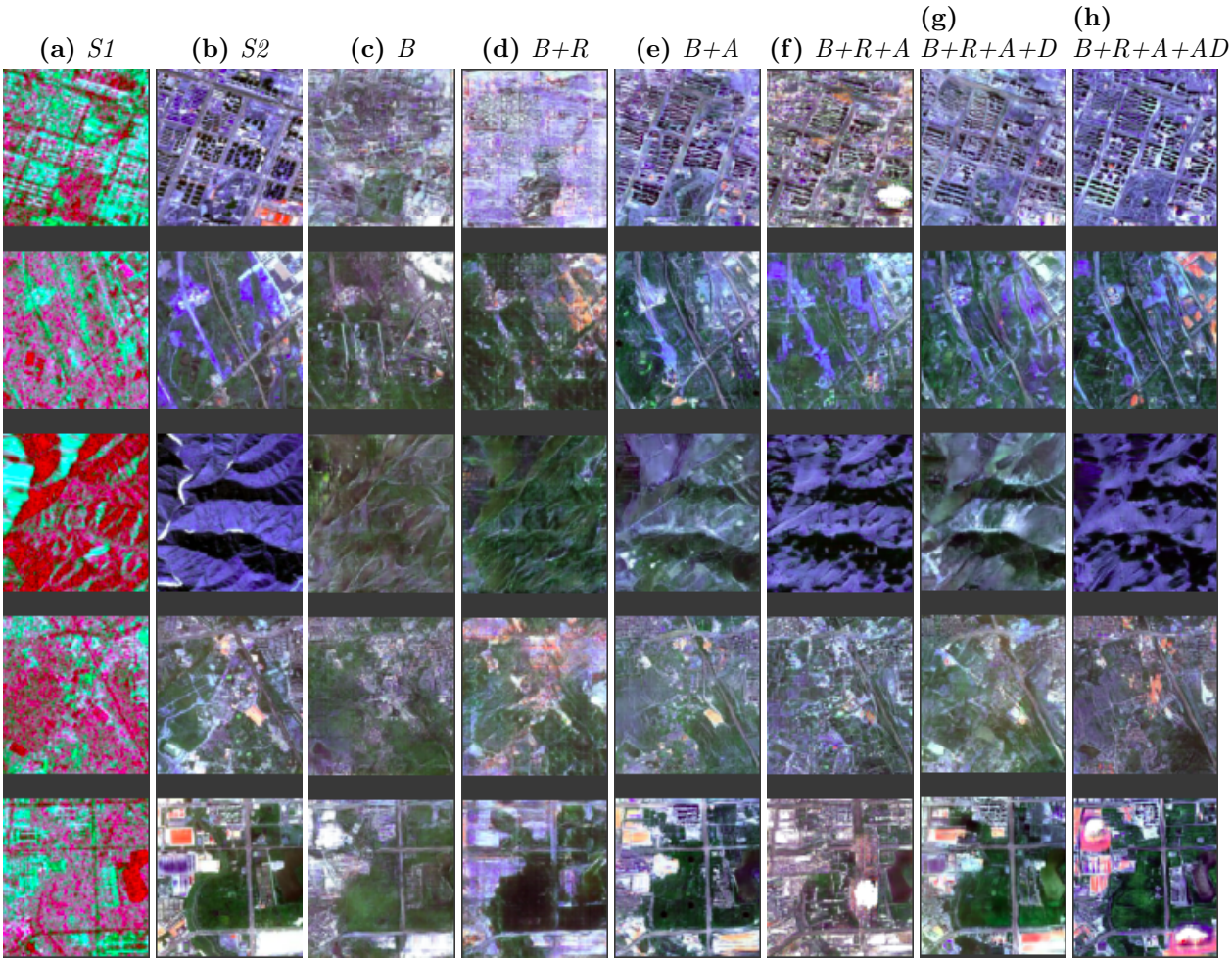


Figure 5.8: *Ablation study visualization. S1 is the processed sentinel-1 image, S2 is the sentinel-2 image, B is the base network, R is a residual connection, A is attention, D is an encoder-decoder based discriminator, and AD is encoder-decoder based discriminator with auxiliary loss.*

5.1.1 Proposed Network

We used an encoder-decoder based network with additive connection in generator and discriminator networks of the conditional GAN as shown in Fig. 5.1. The existing generator architecture used in [165] is improved in this work with the inclusion of residual connections between the blocks and attention modules. The last layer of the generator is 1×1 convolution, which gives three channels of output. We used encoder-decoder based architecture for the discriminator in place of PatchGAN [277] based discriminator. It allowed the discrimination at the image's full resolution and resulted in better discriminative ability. The last layer of the discriminator is 1×1 convolution, resulting in the one-channel feature map at the image's original resolution. We used four encoder and decoder modules in the proposed work, which create features at four different scales. This allows the capturing of scale-invariant details.

5.1.1.1 Encoder Module

The encoder module extracts the multi-scale feature information from the input. We presented the proposed encoder module block diagram in Fig. 5.3. It consists of a sequence of residual blocks, channel, and spatial attention, and a max pooling layer. The residual block contains a sequence of convolution, batch normalization, a pooling layer, and a residual connection. Each encoder module has two outputs, one goes to the next encoder module, and another goes to the corresponding decoder module. We used residual connection as it allowed the alternate path for gradient flow and resolved the vanishing gradient problem. It can be represented as

$$Z(x) = H(x) + x \tag{5.1}$$

Here $Z(x)$ is the residual connection, x is the input, and $H(x)$ is the nonlinear operation skipped by input x.

Attention is the ability to focus on the important locations in the input. We used channel-wise, and spatial attention layers based on CBAM [233] to emphasize important features. The attention module is used in each block of the network. For the input feature

x, channel-wise pooling operation P_c, and spatial pooling operation P_s, the attention can be defined as

$$C_a(x) = conv(P_c(x)) * x \tag{5.2}$$

Here, C_a is channel attention, and *conv* is convolution operation.

$$S_a(C_a) = conv(P_c(C_a)) * C_a \tag{5.3}$$

Here, S_a is spatial attention. The last layer of the encoder contains the pooling operation, which reduces the feature resolution to half.

5.1.1.2 Decoder Module

We used the decoder module as shown in Fig. 5.4. The Transpose convolution layer is used for upsampling the features coming from the previous decoder module. The skip connection features from the corresponding encoder modules are combined using the addition operator followed by a residual block. Lastly, we passed the features to attention layers for further refinement.

5.1.2 Loss Function

We have used a variant of binary cross-entropy loss. Simple binary cross-entropy loss can be written as

$$L(\hat{y}, y) = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \tag{5.4}$$

Here, y is the actual output, and $\hat{y}$ is the produced output. We have generator loss $\mathrm{L^G}$ and discriminator loss $\mathrm{L^D}$. In the above equation for real data x, y = 1 is the real output, and $\hat{y}$ = D(x) is the reconstructed output. For data coming from generator, y = 0 and $\hat{y}$ = D(G(z)). Hence generator and discriminator loss functions can be expressed as

$$\begin{aligned} L_G &= \min[\log(D(x)) + \log(1 - D(G(z))] \\ L_D &= \max[\log(D(x)) + \log(1 - D(G(z))] \end{aligned} \tag{5.5}$$

For a single data point, the combined loss function can be written as

$$L = \min_G \max_D [\log(D(x)) + \log(1 - D(G(z))]$$

Our model uses a conditional GAN: $G : x, z \longrightarrow y$ (where Z: Noise vector, x: Input image and y: Output image) In our case, L_G consists of two parts, L_{cGAN} and L_{L1} where L_{cGAN} is conditional GAN loss and L_{L1} is the mean absolute error which is expressed as Now, Generator loss can be written as

$$\begin{aligned} L_{cGAN}(G, D) &= E_{x,y}[\log D(x, y)] + E_{x,z}[\log(1 - D(x, G(x, z)))] \\ L_{L1}(G) &= E_{x,y,z}[||y - G(x, z)||_1] \\ L_G &= \min_G \max_D [L_{cGAN}(G, D) + \lambda * L_{L1}(G)] \end{aligned} \tag{5.6}$$

where $\lambda = 100$ [278].

Discriminator loss can be expressed as

$$L_d = E_{x,z}[\log D(x, z)] + E_{x,y}[\log(1 - D(x, y))] \tag{5.7}$$

We calculated auxiliary loss for each decoder output of the discriminator. The total loss (L_D) for discriminator is

$$L_D = L_{d1} + 0.2 * L_{d2} + 0.1 * L_{d3} + 0.05 * L_{d4} \tag{5.8}$$

Here, L_{di} is discriminator loss calculated on i^{th} decoder block output.

5.1.3 Dataset and Preprocessing

We used WHU-SEN-City [166] for SAR to RGB translation. It is taken from the European Space Agency, which has two satellites, Sentinel 1 and Sentinel 2. Both satellites can provide multispectral imagery. This dataset consists of 32 Chinese Cities which cover many kinds of geographical scenes such as Mountains, Rivers, Bare lands, etc. The images can be downloaded directly from the ESA website[1]. After taking the dataset from ESA, SAR and RGB images are preprocessed to get Paired SAR-RGB images as shown in figure 5.2. The

[1]ESA Website: `https://scihub.copernicus.eu/dhus/#/home`

preprocessing is required to generate training and testing patches from the dataset. We used the standard approach proposed in the original work [166] for this. The sequence of preprocessing is as follows:

- Reprojection: Used to align the image according to the World geodetic system
- Cropping: To make their size equal according to latitude and longitude
- For Sentinel 2 images:
  - Colour adjustment using R-G-B band
  - Making their color distribution uniform among all the images
- For Sentinel 1 images:
  - Amplitude_VV and Amplitude_VH band are used
  - The third channel is created by computing Amplitude_VH/ Amplitude_VV
  - Minimum and Maximum value is calculated, and the image is linearly transformed into the range of [0, 255]

SNAP tool can be used for Reprojection, Cropping, and various other transformations. Finally, after getting paired SAR and RGB images, these large images are broken down into chunks of 256x256 patches. After pre-processing, we got a large number of 256x256 patches of SAR and RGB images, each with three channels. Sentinel 2 has R-G-B channels, and Sentinel 1 has VH, VV, and VH/VV channels. It can be seen from figure 5.5 that the Sentinel 2 satellite is not able to capture the image properly due to its inability to penetrate through the cloud. In such cases, these predicted RGB images are useful.

5.1.4 Quantitative Parameters

We used peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and chrominance feature-similarity index ($FSIM_c$) for quantitative performance evaluation of

the proposed model. For two images of I_1 and I_2, these parameters are calculated using the following formulas

$$PSNR(I_1, I_2) = 10 \log_{10} \frac{1}{MSE(I_1, I_2)} \tag{5.9}$$

Here, 1 is the maximum pixel value in the images after normalization, and MSE is the mean square root error between I_1 and I_2.

$$SSIM = \frac{(2\mu_{I_1}\mu_{I_2} + c1)(2\sigma_{I_1,I_2} + c2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + c1)(\sigma_{I_1}^2 + \sigma_{I_2}^2 + c2)} \tag{5.10}$$

Here, μ_{I_1} and μ_{I_2} are mean of images I_1 and I_2. It also denotes the luminance of the inputs. The standard deviation of the images are denoted by σ_{I_1}, σ_{I_2}, and σ_{I_1,I_2}. It also represents the contrast of the images.

$$FSIM = \frac{\sum S * PC}{\sum PC} \tag{5.11}$$

Here PC is phase congruency, and S is a similarity between images. The $FSIM_c$ is calculated using an extension of $FSIM$ for the color image as proposed in [279]. The average value of these parameters is used for performance comparison.

5.1.5 Experiments

All the models were trained on the dataset described in Section 4.2.3. The pre-processed paired train set is used to train all the models and evaluated on the test set. We used the dataset split mentioned in the [166]. Adam's optimizer is used for both the generator and discriminator. Random normal initializer with $\mu = 0$ and $\sigma = 0.02$ is used for kernel initializer of convolution operation. We used a batch size of 8. The input images were cropped, resized, and a random jitter was applied before training the models. All the models are implemented in the TensorFlow library.

5.1.6 Results

5.1.6.1 Benchmark comparison

The comparative analysis of the performance with recently published work is presented in Table 5.3. The proposed model has achieved the average values of 12.025,0.382,0.748 for $PSNR$, $SSIM$, and $FSIM_c$ metrics. It is better than the previously published work using the WHU-SEN-City dataset. We improved the average results by 0.822, 0.003, and 0.021 for $PSNR$, $SSIM$, and $FSIM_c$. We achieved a major improvement in the $PSNR$ value and minor improved results for $SSIM$, and $FSIM_c$.

5.1.6.2 City-wise comparison

In Table 5.2, we compared the city-wise result produced by the proposed model with supervised cycle GAN [166] and Pix2Pix [165] models. Our model made better PSNR values for six test city datasets and overall values. The supervised cycle GAN has a better PSNR value for Luoyang and Lanzhou city datasets. For SSIM value, our model has better results for four city datasets. We have better results for seven test city datasets in FSIM_c values.

5.1.7 Qualitative analysis

The proposed model can handle most of the failure cases of the current State-of-the-art (SOTA) method. It can also preserve most of the land-cover information (some examples are shown in figure 5.7. It is preserving bridges structure properly (figure 5.7a, 5.7b, 5.7c). It is also able to translate the cloud-occluded region of the image, can be seen from the figure 5.7g, 5.7h, 5.7i. Some regions of the SAR image could be considered the same using a normal and untrained eye but can be easily distinguished using the predicted RGB image. Due to its cloud-removing property, it can be used in bad weather conditions in which sentinel-2 cannot capture the image adequately. Some of the predicted images and corresponding RGB sentinel-2 images are shown in figure 5.6, which shows the ability of its prediction and structure preservation.

Table 5.1: *Ablation study using PSNR value. Here, B is the base network, R is a residual connection, A is attention, D is an encoder-decoder based discriminator, and AD is an encoder-decoder based discriminator with auxiliary loss.*

Models	luoy	wuh	chan	nan	lanzu	kunm	haer	baod	Avg PSNR
B	8.960	11.020	11.200	11.860	8.430	12.680	11.550	11.270	**10.870**
B + R	6.927	11.289	10.615	14.647	8.503	13.015	10.973	11.123	**11.071**
B + A	7.799	11.601	11.373	14.737	8.606	13.469	11.460	11.970	**11.599**
B + R+ A	7.016	11.882	11.816	14.723	8.259	13.221	11.633	12.446	**11.658**
B + R + A +D	7.946	11.280	11.449	14.334	9.280	13.791	12.030	11.837	**11.749**
B + R +A +AD	7.781	12.898	12.237	15.090	8.468	13.171	11.925	12.476	**12.025**

Table 5.2: *City-wise overall comparison of PSNR, SSIM, and FSIM_c values.*

Metrics	Models	luoy	wuh	chan	nan	lanz	kunm	haer	baod	Avg PSNR
PSNR	Pix2Pix [165]	8.960	11.020	11.200	11.860	8.430	12.680	11.550	11.270	10.870
	Supervised CycleGAN [166]	9.620	11.510	11.480	11.410	9.360	12.930	11.710	11.740	11.203
	Proposed	7.781	12.898	12.237	15.090	8.568	13.171	11.925	12.476	12.025
SSIM	Pix2Pix [165]	0.190	0.333	0.376	0.261	0.189	0.467	0.361	0.379	0.319
	Supervised CycleGAN [166]	0.227	0.399	0.421	0.267	0.294	0.525	0.410	0.472	0.377
	Proposed	0.302	0.393	0.460	0.318	0.276	0.468	0.429	0.408	0.382
FSIM_c	Pix2Pix [165]	0.690	0.729	0.731	0.691	0.670	0.759	0.718	0.741	0.716
	Supervised CycleGAN [166]	0.678	0.744	0.740	0.707	0.678	0.777	0.727	0.764	0.727
	Proposed	0.815	0.700	0.796	0.855	0.772	0.777	0.738	0.766	0.748

5.1.8 Ablation Study

Ablation study for various components of the proposed network is done in the table 5.1. We used pix2pix [165] based conditional GAN as a base network. We achieved the average $PSNR$ value of 10.870. After the integration of channel-wise and spatial attention, it improved to 11.599. We also added a residual block in the base network, which resulted in the PSNR value of 11.071. Using the residual connection with the attention module gives the $PSNR$ value of 11.658. We further improved the discriminator network with encoder-decoder architecture, which resulted in a value of 11.749. Finally, the use of auxiliary loss in the discriminator network value improved to 12.025, which is better than the previous value by 0.85. Visual analysis for the output of the ablation study is presented in Fig. 5.8.

Table 5.3: *Overall comparison of PSNR and SSIM values. Here, the result with the * mark is not comparable because the authors have used different split.*

	Avg. PSNR	Avg. SSIM	Avg. FSIM_c
DualGAN [280]	7.721	-0.062	0.669
CycleGAN [164]	9.715	0.198	0.679
Pix2pix [165]	10.870	0.319	0.716
Sup CycleGAN [166]	11.203	0.377	0.727
Pix2pix + HOPC [281]	11.203	0.379	0.727
Proposed	12.025	0.382	0.748
* FG-GAN [282]	10.93	0.086	-

5.2 MCDNet

High Resolution (HR) satellite images are not easily available. Low-resolution satellites mostly have higher temporal resolution than high-resolution satellites. It makes satellite super-resolution a task important as it makes super-resolved high-resolution data available. In this work, we proposed a novel architecture named MCDNet to handle spatial and temporal pixel dependencies using multiple approaches of global average pooling, multiple-size kernels, and self-attention. The major contributions are as follows:

- We proposed the use of multiple spatial spaces through multiple-size convolution kernels with global average pooling to extract pixel relations for image super-resolution.
- We proposed the use of dot product-based self-attention for pixel relation and to reduce artifacts.
- Dense connection is used in the Dense Residual Attention Module to improve the feature modeling.

5.2.1 Proposed Network

We propose a novel neural network design called MCDNet which takes different spatial-space features into consideration. We do this by using different size kernels for the feature extraction. Figure 5.9 presents a block diagram of MCDNet. Initially, the input is fed to 2D and 3D paths. 3D path extracts the feature information from spatio-temporal dimensions,

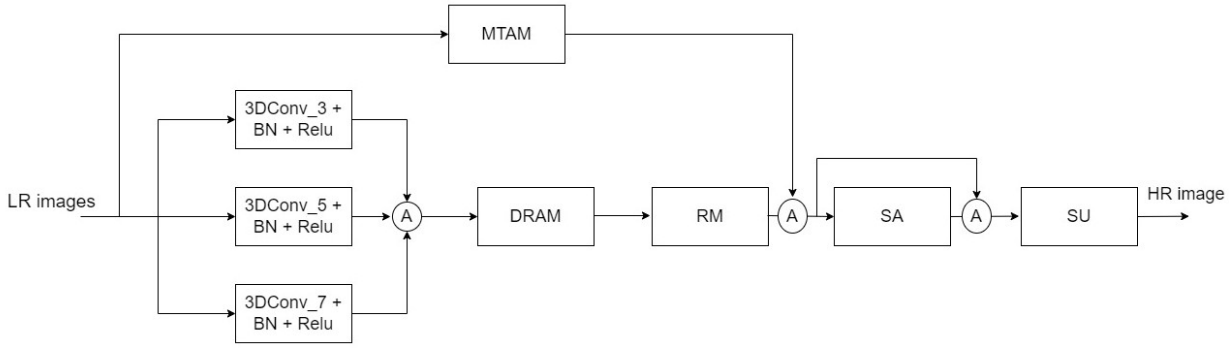


Figure 5.9: *Complete architecture of MCDNet. Here, MTAM is a multi-context temporal attention module, DRAM is a dense connected residual attention module, RM is a reduction module, SA is self-attention, and SU is scaling up. 3DConv_i is 3D convolution with kernel size i, BN is batch normalization, A is addition operation, LR is LR, and HR is high resolution.*

and 2D path processes the spatial information. In the 3D path, input is passed to three parallel streams, each of which extracts features of different the sptial-space. Extracted multi-context features are passed to *Dense Residual Attention Module* (DRAM), which is connected in a dense manner. DRAM produces spatio-temporal feature importance using 3D global pooling. It is followed by temporal reduction, by which the temporal dimension of the features is reduced to one.

In the 2D path, input is passed to three streams which calculate global average pooling on temporal dimension to extract spatial relations. Each stream has a different size kernel to extract multiple spatial-space. Finally, both paths are combined using an addition operation, and overall self-attention maps are calculated. Lastly, features are upsampled using the pixel shuffle operation.

5.2.1.1 Dense Residual Attentive Module

This module takes multi-context input produced using three parallel streams. Figure 5.10 presents a block diagram of DRAM. This module is designed to apply temporal and spatial attention to the input. The attention component of DRAM is first calculated using 3D global average pooling on spatial and temporal dimensions of the input. The pooled features are passed to sigmoid operation and multiplied with input. Multiple 3D convolutions and 3D global average pooling blocks are used in this module. All blocks are connected in a dense

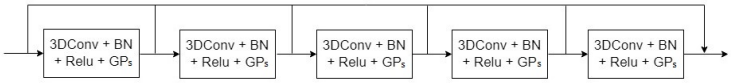


Figure 5.10: *Dense Residual Attention Module. Here, 3DConv_i is 3D convolution, BN is batch normalization operation, GPs is 3D global pooling with sigmoid operation, and connections are combined using the addition operator.*

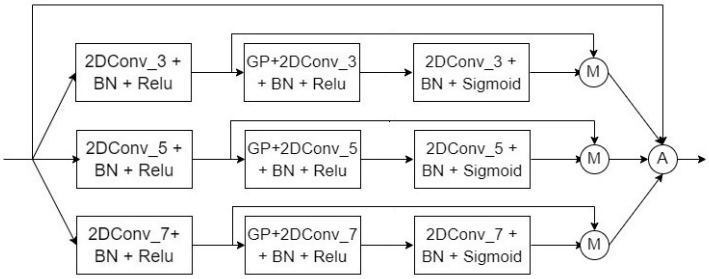


Figure 5.11: *Multi-context Temporal Attention Module. 2DConv_i is 2D convolution with kernel size i, BN is batch normalization, GP is global pooling, M is multiply operation, and A is addition operation.*

manner for better feature representation.

5.2.1.2 Multi context Temporal Attention Module

This module is used to extract the multi-context spatial relation, and the block diagram is shown in Fig. 5.11. We used multiple kernel sizes (3, 5, 7) to generate multi-context features. We used a global pooling operation followed by sigmoid activation to create spatial attention. For each temporal dimension, one attention map is created. These multi-context attentions are combined using residual addition operation.

5.2.1.3 Reduction Module

This module is used as proposed in [2] to decrease the temporal dimension of the features produced from DRAM.

5.2.1.4 Self Attention Module

This module is used to calculate dot product-based self-attention on upsampled features to refine the global spatial patterns. This also helped in the preservation of edges, texture, and finer patterns of images at higher resolutions.

5.2.2 Loss Function

We used the $L1$ loss function in this work. For LR inputs of $[x_1..x_9]$, super-resolved image y_p, and corresponding ground truth image of HR y, it can be represented as

$$L_1 = ||y - y_p||_1 \tag{5.12}$$

5.2.3 Datasets

We evaluate the model on the Proba-V super-resolution dataset. The dataset is provided as part of the PROBA-V SR challenge [91] that took place from November 2018 to May 2019. The satellite clicks pictures with a resolution of 300m every day and pictures with a resolution of 100m every five days of places around the globe. The low-resolution images have a resolution of 128×128 pixels, and HR images have a resolution of 384×384 pixels. Therefore the scaling factor is 3. The images are in a 16-bit format. For each HR image output, nine frames of LR images are passed as input as used in [2, 84, 88, 90, 283].

5.2.4 Experiments

5.2.5 Preprocessing

We used the preprocessing of the dataset as used in RAMS [2]. In this step, low-resolution images are registered, and clean images are selected for training using a minimum threshold value for clearance. Augmentation of selected low-resolution images is done, and image normalization is performed. The input images are cropped and 32×32 pixels patches for low-resolution images, and correspondingly 96×96 pixels patch for high-resolution images is created for model training.

5.2.6 Hyperparamters and Training

We used a batch size of 10, an initial learning rate of 0.0001, and total training epochs of 500. We used NADAM optimizer with values of 0.9 and 0.999 for β_1 and β_2 parameters.

Table 5.4: *Quantitative Performance for SR of NIR band dataset*

	cPSNR	cSSIM
Bicubic	45.44	0.977
Molini [83]	47.21	-
DeepSum (ens) [283]	47.84	-
DeepSum++ [84]	47.93	0.986
RAMS [2]	48.23	0.987
RAMS (ens) [2]	48.51	0.987
TR-MISR [88]	48.54	0.988
PIUNet [90]	48.72	0.988
MCDNet	**49.01**	**0.989**

We implemented this work using the Tensorflow library. This work used a 32 GB graphics card.

5.2.7 Evaluation parameters

For quantitative comparison of the proposed work with the recent approaches [2, 88, 90], we used the corrected variant of peak signal-to-noise ratio ($cPSNR$), and corrected variant of structural similarity index measure ($cSSIM$) metrics as proposed in [197]. The corrected variant of mean square error $cMSE$ is used for the calculation of $cPSNR$. The $cMSE$, $cPSNR$ and $cSSIM$ are calculated using the following equations

$$cMSE = \frac{||x^{HR(p,q)} * c^{HR(p,q)} - x^{SR(p,q)} * c^{SR(p,q)} - b * c^{SR(p,q)}||_2^2}{||c^{SR(p,q)}||_1} \tag{5.13}$$

Here, HR is high resolution image, x^{SR} is a super-resolved image, c is a clearance mask, and b is brightness deviation.

$$cPSNR = max_{p,q\in[0,6]} 10 \log_{10} \frac{(2^{16}-1)^2}{cMSE} \tag{5.14}$$

Here, $(2^{16}-1)$ is the maximum value for a 16 bit input.

$$cSSIM = max_{p,qin[0,6]} SSIM(x^{HR(p,q)} * c^{HR(p,q)}, x^{SR(p,q)} * c^{SR(p,q)} + b) \tag{5.15}$$

Table 5.5: *Quantitative Performance for SR of RED band dataset*

	cPSNR	cSSIM
Bicubic	47.34	0.984
Molini [83]	49.52	-
DeepSum (ens) [283]	50.0	-
DeepSum++ [84]	50.08	0.991
RAMS [2]	50.17	0.991
RAMS (ens) [2]	50.44	0.991
PIUNet [90]	50.62	0.992
TR-MISR [88]	50.67	0.992
MCDNet	**50.72**	**0.993**

Table 5.6: *Ablation study using NIR dataset*

	cPSNR	cSSIM
B	48.23	0.987
B + MC	48.52	0.987
B + MC + D	48.67	0.987
B + MC + D + SA	49.01	0.989

5.2.8 Analysis

The quantitative comparisons are presented in Table 5.4 for multi-frame super-resolution of the NIR band. The traditional BiCubic based approach achieved cPSNR and cSSIM values of 45.44 and 0.977. The results of cPSNR are improved to 47.21 and 47.84 by Molini et al. [83], and DeepSum [283]. DeepSum ++ [84] achieved the cPSNR and cSSIM values of 47.93 and 0.986, which is further enhanced to 48.51 and 0.987 by RAMS [2]. TR-MISR [88] get the values of 48.54 and 0.988 which are improved to 48.72 and 0.988 by PIUNet [90] cPSNR and cSSIM. Our method achieved $cPSNR$ and $cSSIM$ values of 49.01 and 0.989 with an improvement of 0.29% in cPSNR values and 0.01% in cSSIM.

In Table 5.5, results for RED band super-resolution are shown. BiCubic approach results in cPSNR and cSSIM values of 47.34 and 0.984. PSNR value is further improved by Molini et al. [83] to 49.52 and 50 by DeepSum [283]. DeepSum++ [84] achieved cPSNR and cSSIM values of 50.08, and cPSNR is improved to 50.44 by RAMS [2]. PIUNet [90] enhanced

Table 5.7: *Ablation study using RED dataset*

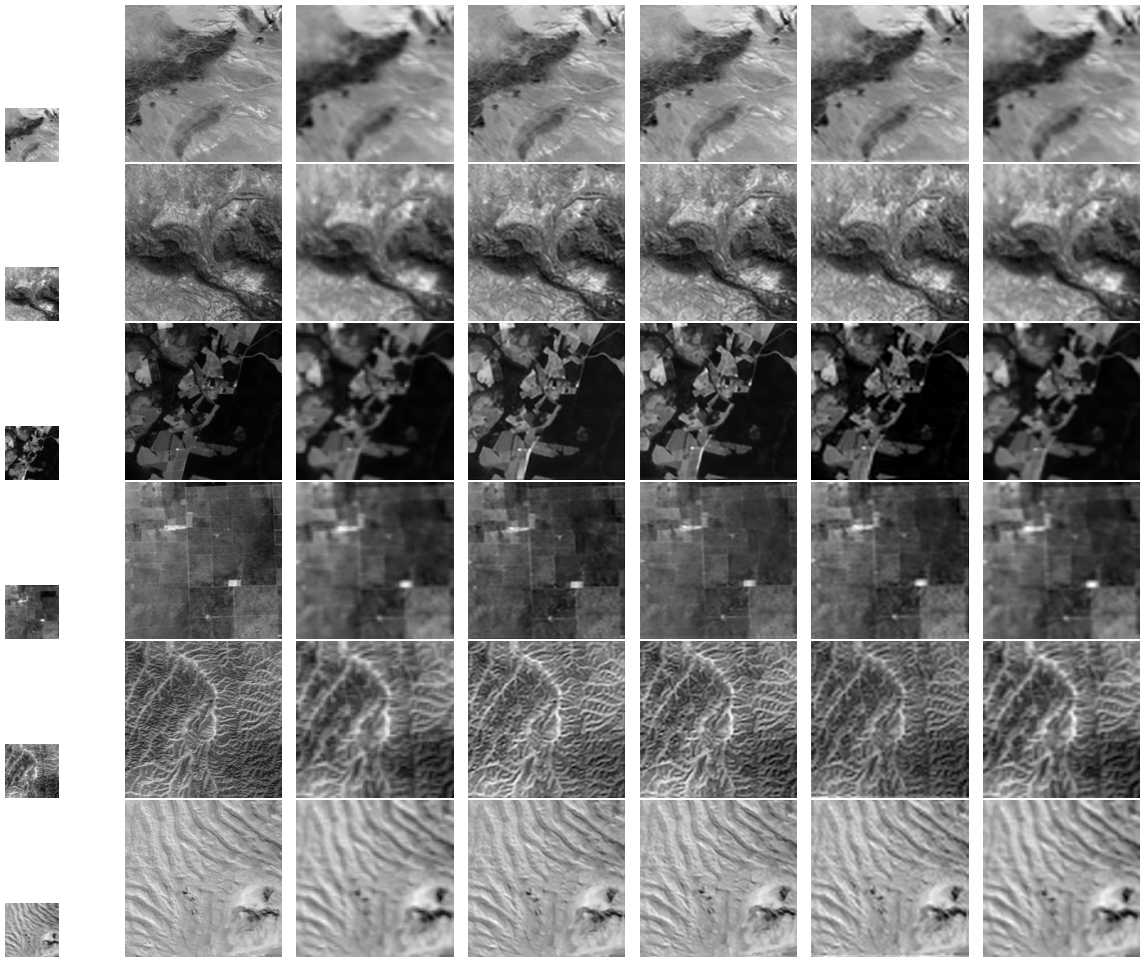
	cPSNR	cSSIM
B	49.08	0.989
B + MC	49.57	0.989
B + MC + D	49.74	0.991
B + MC + D + SA	50.72	0.993

cPSNR and cSSIM values to 50.62 and 0.992, and cPSNR value improved to 50.67 by TR-MISR [88]. Our method achieved cPSNR and cSSIM values of 50.72 and 0.993 with 0.05% and 0.01% improvement in cPSNR and cSSIM values.

Qualitative comparison of output produced by various methods [2, 88, 90] on RED and NIR datasets are shown in Fig. 5.12 and Fig. 5.13. The output produced by our methods is more consistent with the high-resolution ground truth and better compared to other methods.

5.2.9 Ablation Study

In Table 5.6 and Table 5.7, ablation studies on NIR and RED datasets are presented to highlight the importance of each module of the MCDNet. We started with a baseline network based on RAMS [2] with minor variations. After training, the baseline network achieves cPSNR and cSSIM values of (48.23, 0.987) and (49.08, 0.989) for NIR and RED band datasets. The inclusion of MTAM improved the value of cPSNR for both datasets. Adding DRAM further enhanced the cPSNR values and improved the cSSIM. Incorporating self-attention further improved the values of cPSNR and cSSIM to (49.01, 0.989) and (50.72, 0.993) for NIR and RED datasets. The highest improvements in cPSNR and cSSIM values are achieved by the inclusion of self-attention, and MTAM and DRAM attain the next highest improvement. All three modules produce the multi-context representations, which improved the results.

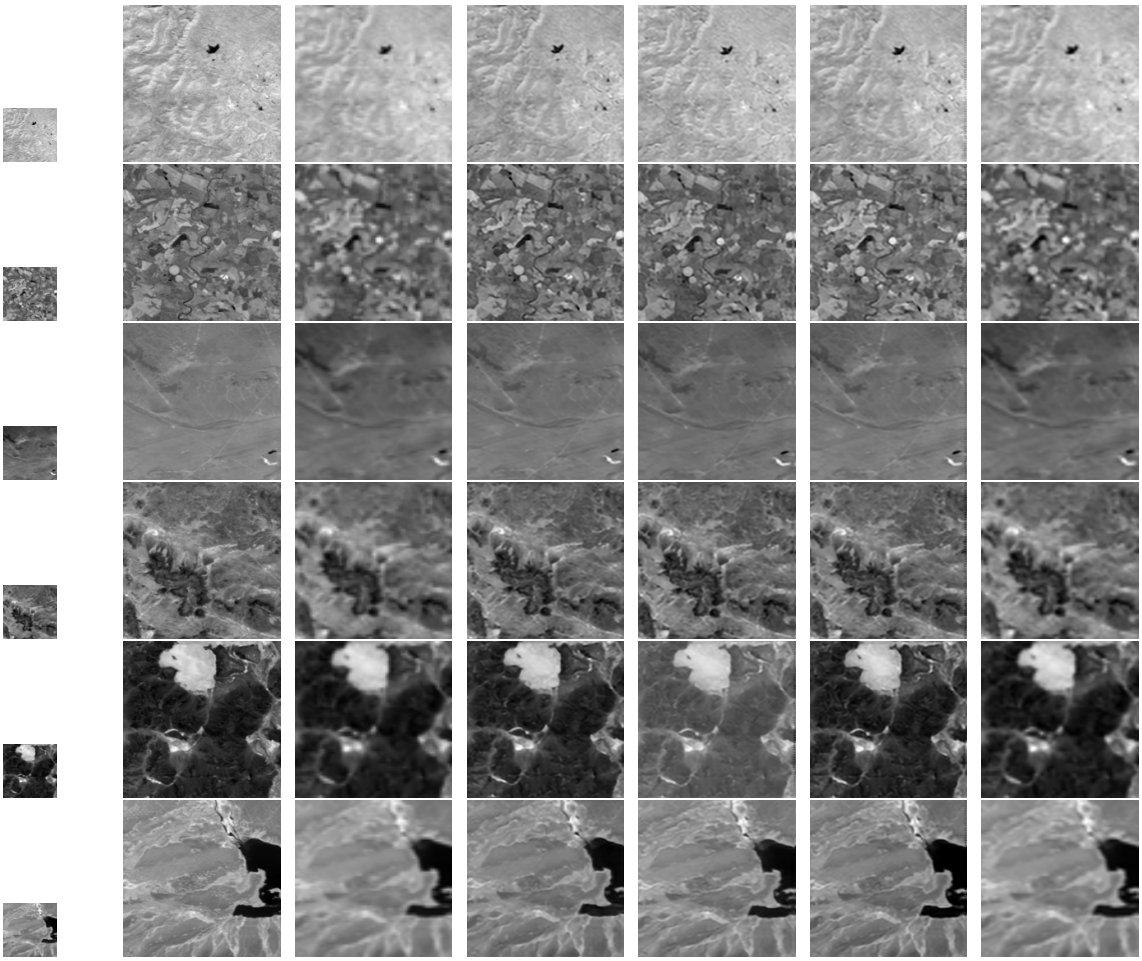


(a) *L* (b) *H* (c) *PIUNet* (d) *RAMS* (e) *TR-MISR* (f) *MCDNet* (g) *Bicubic*

Figure 5.12: *Visual comparison of SR of RED band patches of input size* 64×64 *pixels and output size of* 192×192 *pixels. Here, L is one of LR inputs shown, H is the ground truth high-resolution image, and the output of PIUNet, RAMS, TR-MISR, MCDNet, and Bicubic in column (a) to (g).*

5.3 Summary

In this work, we proposed the improvement in conditional GAN for SAR to optical translation with the use of multi-scale attention enhanced generator and discriminator with auxiliary loss. The encoder-decoder architecture allowed the learning of geometrically rich image generation and helped in better discrimination ability for the discriminator. The utilized residual and skip connections mitigated the vanishing gradient problem and helped in multi-scale feature sharing. We demonstrated a comparative study on one benchmark dataset and achieved the SOTA result for SAR to optical translation for three quantitative parameters. We presented a detailed survey on the recent work, and a detailed ablation study for various



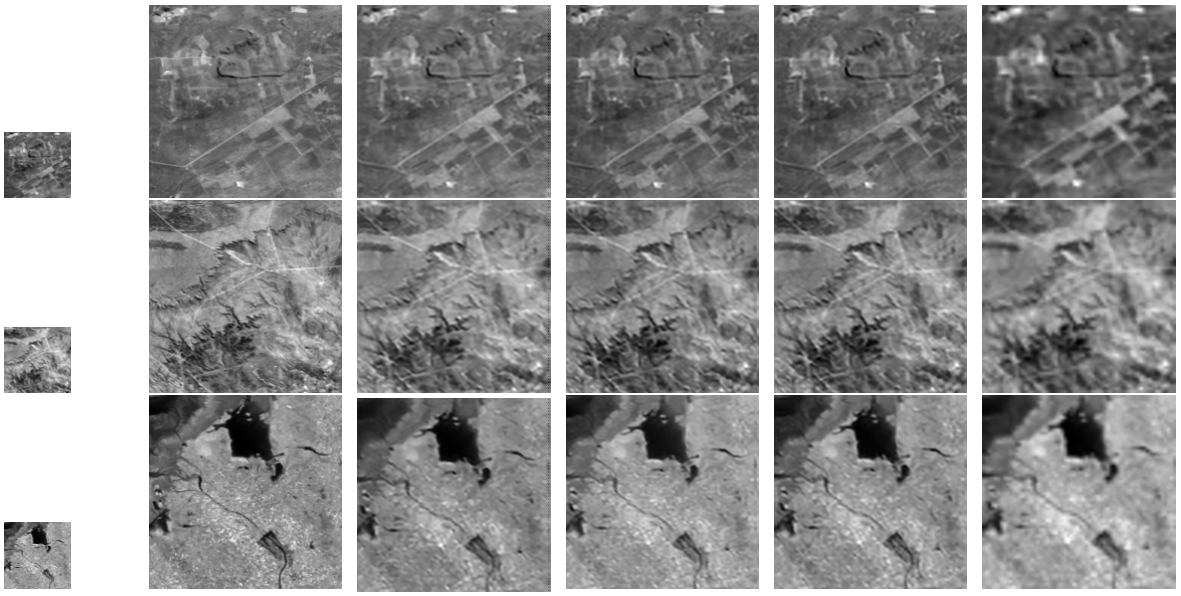
(a) *L* **(b)** *H* **(c)** *PIUNet* **(d)** *RAMS* **(e)** *TR-MISR* **(f)** *MCDNet* **(g)** *Bicubic*

Figure 5.13: *Visual comparison of super-resolution of NIR band patches of input size* 64×64 *pixels and output size of* 192×192 *pixels. Here, L is one of LR inputs shown, H is the ground truth HR image, and the output of PIUNet, RAMS, TR-MISR, MCDNet, and Bicubic in column (a) to (g).*

components' usefulness was conducted with analysis.

This work presents a multiple context-based network named MCDNet, which is effective for multi-frame image super-resolution of satellite images. The proposed approach has used global average pooling, different size kernels, and dot product-based self-attention to capture spatial and temporal relations effectively. The proposed MCDNet is validated through quantitative and qualitative experiment results. The results reveal that it has achieved SOTA results on the PROBA-V benchmark dataset on cPSNR and cSSIM metrics. Ablation studies are presented to verify the significance of each module in the MCDNet.

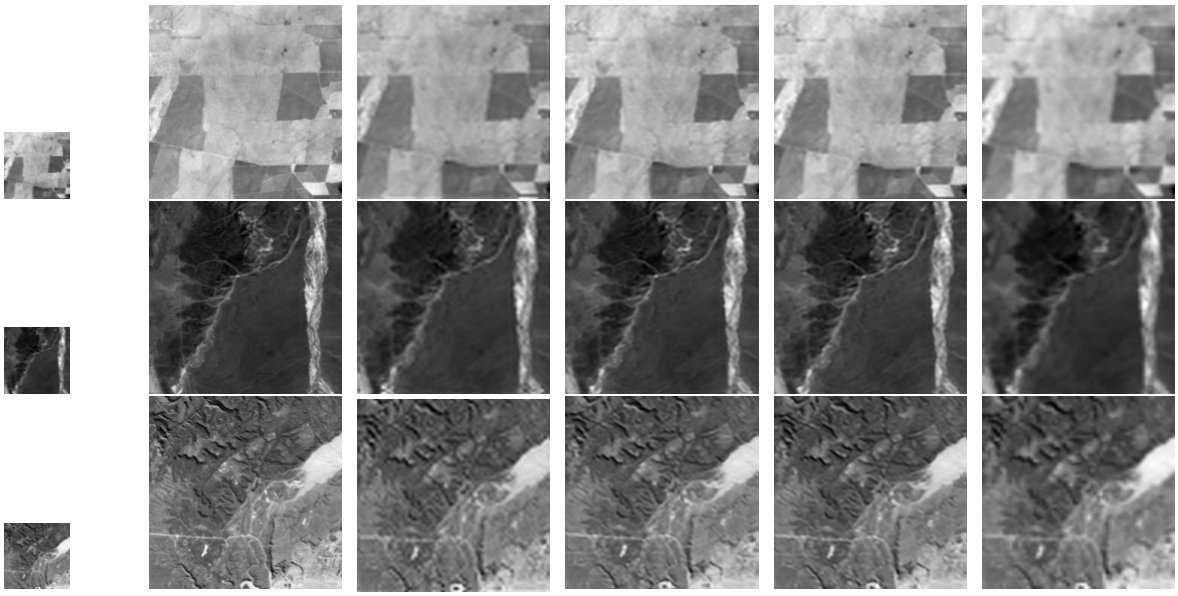
The next contributory chapter presents the details of proposed models for UDA task. It



(a) *L* **(b)** *H* **(c)** *B+MC+D+SA* **(d)** *B+MC+D* **(e)** *B+MC* **(f)** *B*

Figure 5.14: *Visual comparison of SR of NIR band patches for ablation study. Here, L is one of the low-resolution inputs, H is the ground truth HR image, B is the base network, MC is the multi-context temporal attention module, D is the dense residual attentive module, and SA is self-attention.*

is an important RSVT which allows the transfer of knowledge acquired from labeled source domain to unlabelled target domain.



(a) *L* **(b)** *H* **(c)** *B+MC+D+SA* **(d)** *B+MC+D* **(e)** *B+MC* **(f)** *B*

Figure 5.15: *Visual comparison of super-resolution of RED for ablation study. Here, L is one of the low-resolution inputs, H is the ground truth HR image, B is the base network, MC is the multi-context temporal attention module, D is the dense residual attentive module, and SA is self-attention.*

'As you start to walk on the way, the way appears."

~Rumi

6

Reconstruction assisted unsupervised domain adaptive techniques

In this chapter, we discuss the fourth contribution to the Unsupervised Domain Adaptation (UDA) task. UDA involves adaptation of knowledge learned from labeled source domain to unlabelled target domain. The source and target domain may have differences in modality (sensors), capturing conditions, resolutions, etc. UDA for Semantic Segmentation (SS) is an important Remote Sensing Vision Task (RSVT) because creating precise labels for new target domain is difficult. UDA for nighttime Object Tracking (OT) is another important RSVT which enables tracking of object at low-light night conditions. We proposed two Deep learning (DL) models named Masked Domain Adversarial Adaptation Network (MDAANet)

and Reconstruction Assisted Domain Adaptation (RADA) for domain adaptive SS and OT. The significant contributions of this chapter are summarized as follows:

- The proposed MDAANet model incorporates masked domain dual adaptation, joint adversarial adaptation, domain reconstruction consistency, and feature dissimilarity alignment for UDA in SS.
- The proposed RADA framework utilizes static style transfer and multi-level adaptation for UDA in OT.

The rest of the chapter is organized as follows. In Section 6.1 the details of the proposed MDAANet architecture, experiments, and results are shown. Section 6.2 described the proposed model named RADA and experimentation and results details. The chapter is concluded in Section 5.3.

6.1 MDAANet

UDA tasks can be implemented with alignment at input, feature, and output levels. Some methods [96,98] used multimodal inputs for multi-source domain alignment. Self-training is an approach to generate pseudo labels of target domain during the training where the model used source domain data and it is utilized by several methods [96,101–103,105–107,140,199]. Consistency based approach [107–111, 124] enforces consistency constraints on outputs of different augmented inputs. The nonadversarial approach has been used by some methods [112–114]. Correlation between features is proposed in [114] for effective alignment. KL divergence at outputs and features levels are used for adaptation in [106,115]. Output level adaptation is proposed in [54, 106, 116, 118, 119, 200] which aligns the category outputs of domains to reduce domain shift. Class centroids are calculated for output level adaptation in [96, 106, 113] to reduce class distribution shift. The feature level alignment is utilized in [96,106,118,120,130,137,200] to attain domain invariant features. Differences in resolution, sensors, capturing conditions, and inter-class relations between domains are common in remote sensing datasets, and these need to be handled effectively.

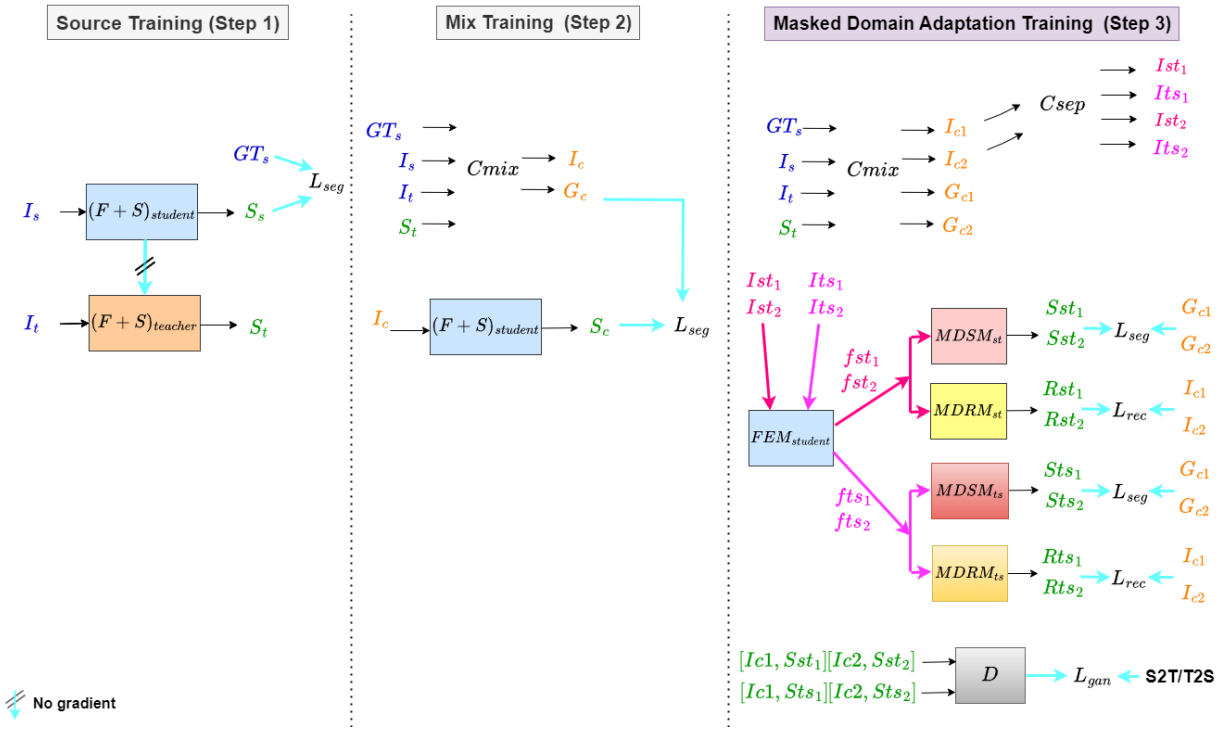


Figure 6.1: *One sequence of MDAANet training. Here, I_s, I_t, and GT_s are the input source image, input target image, and input segmentation ground truth. I_c, G_c, Ist, Its, R_d, and S_d are mixed input, mixed ground truth, domain mask input with source data, domain mask input with target data, reconstructed input, and predicted output for domain d. $F+S$, FEM, $MDSM$, $MDRM$, and D are Feature Extraction Module with Segmentation Module, Feature Extraction Module, Masked Domain Segmentation Module, Masked Domain Reconstruction Module, and Masked Domain Discriminator Module. L_{seg} and L_{rec} are segmentation and reconstruction losses. $S2T$, and $T2S$ are source-to-target and target-to-source domain learning.*

In this work, we proposed a MDAANet with a masked domain dual adaptation approach, joint adversarial alignment, consistency enforcement, and feature dissimilarity-based alignment for effective UDA. The proposed approach allows adaptation at input, feature, and output levels, producing better UDA for semantic segmentation. It increased the inter-class dissimilarities and attained input-output level combined adaptation. The major contributions are as follows:

1. Masked Domain Dual Adaptation approach is proposed for effective domain alignment at input, feature, and output levels using MDRM, MDSM, and common FEM.

2. Joint Adversarial Alignment attains the combined input-output level adaptation, resulting in better segmentation results using MDDM.

3. Domain Reconstruction Consistency (DRC) and Restructure Consistency (RC) are proposed to align domain outputs and generate input-consistent results.

4. Feature Dissimilarity based Alignment (FDA) is proposed to increase the inter-class distance for effective SS.

Algorithm 2: MDAANet workflow pseudocode. Here, I_s, I_t, S_s, and S_t are the source image, target image, source ground truth, and target predicted label.

Data: $(I_s,\ I_t,\ S_s) \in Data_s \times Data_t$
Result: S_t
while *!Null*$(I_s,\ I_t)$ **do**
  if *training* **then**
    $FEM_{tea}, SM_{stu} \leftarrow ema(FEM_{tea}, SM_{stu})$;
    $P_s = SM_{stu}(FEM_{stu}(I_s))$;
    $L_s = Loss(S_s, P_s)$;
    $BackpropogateL_s$;
    $L_t = SM_{tea}(FEM_{tea}(I_t))$;
    $I_{cm}, L_{cm} = Cmix(I_s, I_t, L_s, L_t)$;
    $P_{cm} = SM_{stu}(FEM_{stu}(I_{cm}))$;
    $L_{cm} = Loss_{seg}(S_{cm}, P_{cm})$;
    $BackpropogateL_{cm}$;
    $I_{c1}, I_{c2}, G_{c1}, G_{c2} = Cmix(I_s, I_t, L_s, L_t)$;
    $I_{st1}, I_{ts1}, I_{st2}, I_{ts2} = Csep(I_{c1}, I_{c2})$;
    for $k \leftarrow 1$ *to* 2 **do**
      $f_{st_k}, f_{ts_k} = FEM_{stu}(I_{st_k}, I_{ts_k})$;
      $P_{st_k} = MDSM_{st}(f_{st_k})$;
      $P_{ts_k} = MDSM_{ts}(f_{ts_k})$;
      $R_{st_k} = MDRM_{st}(f_{st_k})$;
      $R_{ts_k} = MDRM_{ts}(f_{ts_k})$;
      $L_{seg} = Loss(P_{stk}, G_{c1}) + Loss(P_{tsk}, G_{c2})$;
      $L_{rec} = Loss(R_{stk}, I_{c1}) + Loss(R_{tsk}, I_{c2})$;
    $L_{cons} = Loss(P_{st1}, P_{st2}, P_{ts1}, P_{ts2}, R_{st1}, R_{st2}, R_{ts1}, R_{ts2})$;
    $L_{dism} = Loss(f_{st1}, f_{st2})$;
    $BackpropogateL_{seg}, L_{rec}, L_{cons}, L_{dism}$;
    $L_{gan} = Loss(Disc(P_{st1}, I_{c1}, P_{ts1}, P_{st2}, I_{c2}, P_{ts2}))$;
    $BackpropogateL_{gan}$;
  else if *testing* **then**
    $P_t = SM_{stu}(FEM_{stu}(I_t))$;
    $returnP_t$;

6.1.1 Proposed Framework

The Proposed framework consists of the Mask Domain Dual Adaptation approach (MDDA), Joint Adversarial Alignment (JDA), Consistency Enforcement (CE), and Feature Dissimilarity Alignment (FDA) approaches, which MDAANet utilizes for effective unsupervised domain adaptative semantic segmentation. Fig. 6.1 represents the proposed framework training sequence, which consists of three-step training that uses source data for the first step and source and target data in the next two steps. The gradients are calculated and backpropagated in all steps separately. Two networks (student and teacher) are used in the first step, and one network (student) is utilized in the next two steps. The ClassMix [284] strategy generates a combined representation of source and target domains, which are used in the second and third steps. In the second step of the training sequence student network is trained on mixed inputs with a common Feature extractor module (FEM) and Segmentation Module (SM). In the third step, the proposed masked domain reconstruction modules (MDRM) and masked domain segmentation modules (MDSM) are used along with the common Feature extractor Module (FEM) used in the student network. The different components of the proposed MDAANet are depicted in Fig. 6.2 and Fig. 6.3. The pseudocode of the MDAANet workflow is shown in Algorithm 2. MDAANet achieved SOTA results for five domain adaptation tasks for aerial semantic segmentation.

6.1.2 Mask Domain Dual Adaptation approach

We proposed a Masked Domain Dual Adaptation (MDDA) approach for effective domain adaptive semantic segmentation. It used masked domain inputs to generate complete mixed inputs and mixed segmentation maps to attain all levels of adaptation. The source and target inputs are combined using ClassMix [284]. With source domain image I_s and target domain image I_t, the ClassMix operation can be defined as

$$\begin{aligned} I_{c1} &= M * I_s + (1 - M) * I_t \\ I_{c2} &= (1 - M) * I_s + M * I_t \end{aligned} \tag{6.1}$$

Here, I_{c1}, I_{c2} are the dual mix images, and M is the class masks having the values of 1 for the randomly selected few classes in source data. In our proposed work, we used both domain images and generated four masked domain inputs $I_{1_{S\to T}}$, $I_{1_{T\to S}}$,$I_{2_{S\to T}}$, $I_{2_{T\to S}}$

$$\begin{aligned} I_{1_{S\to T}} &= M * I_s \\ I_{1_{T\to S}} &= (1-M) * I_t \\ I_{2_{S\to T}} &= (1-M) * I_s \\ I_{2_{T\to S}} &= M * I_t \end{aligned} \tag{6.2}$$

These masked domain images are passed to Masked Domain Reconstruction Modules ($MDRM_{S\to T}$, $MDRM_{T\to S}$) and Masked Domain Segmentation Modules ($MDSM_{S\to T}$, $MDSM_{T\to S}$) respectively to reconstruct complete mixed inputs and to generate complete mixed segmentation maps. These masked domain images are related, and the original images can be recreated using the following relation.

$$\begin{aligned} I_S &= I_{1_{S\to T}} + I_{2_{S\to T}} \\ I_T &= I_{1_{T\to S}} + I_{2_{T\to S}} \end{aligned} \tag{6.3}$$

This property is used in input-output alignment using Restruture Consistency (RC) Enforcement to produce more consistent input-output pairs.

6.1.3 Joint Adversarial Alignment

We proposed Joint Adversarial Alignment (JAA), which is implemented using MDDM. Our novel idea is to utilize input-output pairs for adversarial training, which helps in joint alignment at input and output levels, which is missing in the earlier approaches. JAA aligns reconstructed input-output pairs produced by ($MDSM_{S\to T}$, I) and ($MDSM_{T\to S}$, I). JAA helps in better bonding between the input image and output segmentation map and ensures consistent input-output reconstruction in both directions. The loss function for JAA is calculated using the following equations.

$$L_{g_2}(G,D) = E_{mi,st}[\log D(mi,st)] + E_{mi,ts}[\log(1 - D(mi,ts)))]$$
$$L_{adv2} = \min_{G} \max_{D}[L_{g_2}(G,D)] \tag{6.4}$$

Here, mi is mixed input commonly used in both direction adaptation. Predicted outputs $(S_{S\to T}, S_{T\to S})$ of both domains as represented as st and ts. G is the MDAANet network used in the third step, and D is the Masked Domain Discriminator Module (MDDM).

6.1.4 Consistency Enforcement

We proposed Domain Reconstruction Consistency (DRC) and Restructure Consistency (RC) to generate consistent features. It is implemented using consistency loss (L_{cons}), the combination of losses calculated for DRC and RC.

6.1.4.1 Domain Reconstruction Consistency (DRC)

DRC is enforced using loss based consistency for reconstruted input and output pairs, which are produced by $MDSM_{S\to T}$, $MDRM_{S\to T}$, $MDSM_{T\to S}$, and $MDRM_{T\to S}$. DRC makes sure that outputs are consistent with each other in both directions of generation (source to target or target to source). SSIM loss ($Lrdrc_{ssim}$) is calculated using generated outputs of $MDRM_{S\to T}$ and $MDRM_{T\to S}$ for input level consistency. Dice loss ($Lsdrc_{dc}$) is used to align the outputs generated by $MDSM_{S\to T}$ and $MDSM_{T\to S}$. The DRC loss (L_{cons1}) is calculated as

$$L_{cons1} = Lsdrc_{dc} + Lrdrc_{ssim} \tag{6.5}$$

6.1.4.2 Restructure Consistency (RC)

RC is implemented using consistency loss, which is calculated between restructured outputs and original domain inputs. The restructured output is generated from outputs of $MDSM_{S\to T}$, $MDSM_{T\to S}$, $MDRMS \to T$, and $MDSM_{T\to S}$ following the equation 2. RC put additional conditions on masked reconstruction and segmentation modules to generate

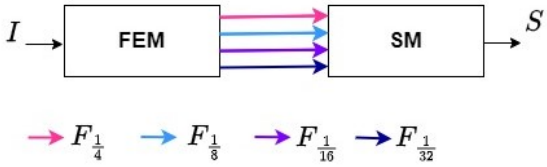


Figure 6.2: *Different components of the MDAANet during step 1 and step 2 of training. Here, FEM is Feature Extraction Module, and SM is Segmentation Module. F is the features extracted on different scales. I, S is the input image and predicted segmentation map.*

outputs consistent with the original domain inputs. The RC loss (L_{cons2}) is calculated as

$$L_{cons2} = Lsrc_{dc} + Lrrc_{ssim} \tag{6.6}$$

Here, $Lsrc_{dc}$ is calculated using dice loss between the original domain segmentation ground truths and restructured predicted segmentation outputs. $Lrrc_{dc}$ is calculated using SSIM loss between original domain inputs and predicted restructured reconstruction outputs. The total consistency loss is calculated as follows:

$$L_{cons} = L_{cons1} + L_{cons2} \tag{6.7}$$

6.1.5 Feature Dissimilarity based alignment

We proposed Feature Dissimilarity based Alignment (FDA) to increase the distance between non-similar feature groups of the source domain. The feature distance between $F_{1_{S\to T}}$ and $F_{2_{S\to T}}$ are extracted using FEN after passing the mix maksed inputs $I_{1_{S\to T}}$ and $I_{2_{T\to S}}$. As these mixed masked inputs represent different classes, the extracted feature group using them should be far from each other. FDA achieved maximization of non-similar feature distance using the minimization of cosine similarity between $F_{1_{S\to T}}$ and $F_{2_{T\to S}}$. FDA helps inter-class feature-level separation in source domains, which is important for effective semantic segmentation. The feature dissimilarity loss is represented as L_{dis}, which is added to the total loss.

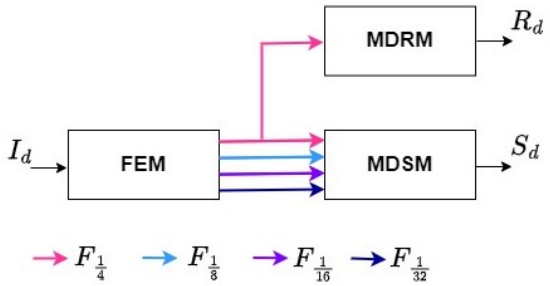


Figure 6.3: *Different components of the MDAANet in the third step of the training. Here, FEM is Feature Extraction Module, MDRM is Masked Domain Reconstruction Module, and MDSM is Masked Domain Segmentation Module. I_d, R_d, and S_d are masked domain input, reconstructed mask domain input, and predicted mask domain segmentation output.*

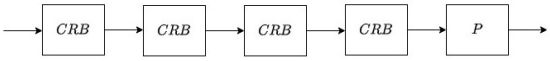


Figure 6.4: *Different components of the Masked Domain Reconstruction Module. Here, CRB is convolution with batch normalization and leaky relu operation, and P is subpixel convolution operation.*

6.1.6 MDAANet components

MDAANet consists of a common Feature Extraction Module (FEM), Segmentation Module (SM), Masked Domain Reconstruction Module (MDRM), Masked Domain Segmentation Module (MDSM) and Masked Domain Discriminator Module (MDDM). MDAANet used Joint Adversarial Alignment (JAA) to align jointly at input-output levels. MDAANet utilized Consistency Enforcement and Feature Dissimilarity based alignment to achieve enhanced domain adaptive segmentation outputs.

6.1.6.1 Feature Extractor Module

Feature Extractor Module (FEM) is used as a backbone to extract the features from the input. Transformer based encoder is utilized as proposed in [285] in FEM. It produced the output features in the form of feature maps ($F_{\frac{1}{4}}, F_{\frac{1}{8}}, F_{\frac{1}{16}}, F_{\frac{1}{32}}$) with resolutions of $\frac{1}{4}^{th}$, $\frac{1}{8}^{th}$, $\frac{1}{16}^{th}$, $\frac{1}{32}^{th}$ of the input resolution. All feature maps ($F_{\frac{1}{4}}, F_{\frac{1}{8}}, F_{\frac{1}{16}}, F_{\frac{1}{32}}$) are passed to the Segmentation Module (SM) in the first two steps of training and to Masked Domain Segmentation Modules (MDSM) in the third step. The highest resolution feature map $F_{\frac{1}{4}}$)

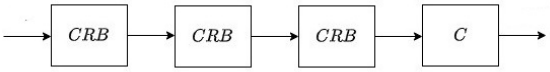


Figure 6.5: *Different components of the Masked Domain Discriminator Module. Here, CRB is convolution with batch normalization and leaky relu operation, and C is class convolution operation.*

is also passed to the Masked Domain Reconstruction Module (MDRM) to produce mixed reconstruction output in the third step. For source input I_s, mixed input I_c, the FEM operation can be represented as

$$\begin{aligned} F_s &= FEM(I_s) \\ F_c &= FEM(I_c) \\ F_{S\to T} &= FEM(I_{S\to T}) \\ F_{T\to S} &= FEM(I_{T\to S}) \end{aligned} \tag{6.8}$$

Here, F is extracted features by FEM.

6.1.6.2 Segmentation Module

Segmentation Network (SM) takes four feature maps ($F_{\frac{1}{4}}, F_{\frac{1}{8}}, F_{\frac{1}{16}}, F_{\frac{1}{32}}$) from the FEM and produces the segmentation map S. We used MLP based network as proposed in [285] for the Segmentation Network.

6.1.6.3 Masked Domain Reconstruction Module

The Masked Domain Reconstruction Module (MDRM) is proposed to reconstruct the complete class-mixed inputs from maksed domain inputs. MDRM components are depicted in Fig. 6.4. MDRM also uses common features extracted by FEM. MDRM helps to reconstruct one domain's input using another domain's features. In this way, it enables the input level domain adaptation for the missing class mask. We used two MDRMs to achieve the both-way input level alignment between the source and target domain. For mask domain inputs $I_{S\to T}, I_{T\to S}$ and extracted highest resolution features $F_{S\to T\frac{1}{4}}, F_{T\to S\frac{1}{4}}$, third step $MDRM$

outputs can be represented as

$$R_{S\to T} = MDRM_{S\to T}(F_{S\to T\frac{1}{4}})$$
$$R_{T\to S} = MDRM_{T\to S}(F_{T\to S\frac{1}{4}}) \tag{6.9}$$

6.1.6.4 Masked Domain Segmentation Module

Masked Domain Segmentation Module (MDSM) is proposed to produce mixed segmentation masks from masked domain inputs. MDSM uses common features extracted by FEN. Two MDSM learns to generate mixed segmentation mask that includes information on other domain segmentation mask using the input of another domain. Two MDSMs are used to achieve alignment in both directions (source to target and target to source). MDSM allows output level alignment between source and target domains. For mask domain inputs $I_{S\to T}, I_{T\to S}$ and extracted features $(F_{S\to T\frac{1}{4}}, F_{S\to T\frac{1}{8}}, F_{S\to T\frac{1}{16}}, F_{S\to T\frac{1}{32}}, F_{T\to S\frac{1}{4}}, F_{T\to S\frac{1}{8}}, F_{T\to S\frac{1}{16}}, F_{T\to S\frac{1}{32}}$, third step $MDSM$ outputs can be represented as

$$S_{S\to T} = MDSM_{S\to T}(F_{S\to T\frac{1}{4}}, F_{S\to T\frac{1}{8}}, F_{S\to T\frac{1}{16}}, F_{S\to T\frac{1}{32}})$$
$$S_{T\to S} = MDSM_{T\to S}(F_{T\to S\frac{1}{4}}, F_{T\to S\frac{1}{8}}, F_{T\to S\frac{1}{16}}, F_{T\to S\frac{1}{32}}) \tag{6.10}$$

6.1.6.5 Masked Domain Discriminator Module

Masked Domain Discriminator Module (MDDM) is proposed for the combined input-output alignment. The concatenated input-output pair is passed to the discriminator. It consists of convolution, leaky relu, and batch normalization operations, as shown in Fig. 6.5. The final layer is the classification layer to differentiate the reconstructed input-output pairs. MDDM is used to align between input-output pairs $([S_{S\to T}, I], [S_{T\to S}, I])$ generated by $(MDSM_{S\to T})$ and $(MDSM_{T\to S})$.

6.1.7 Training Sequence

Student-Teacher based learning is applied to generate the pseudo labels for target images. The student network is trained on source images in the first steps and source and target images in the next steps. The exponential moving average of the weight of the student

network is used to update the weight of the teacher network. The teacher's weights are updated, and pseudo-target labels are generated after each sequence (three steps). The weight update can be represented as

$$w_t{}^i = \lambda w_s{}^{i-1} + (1-\lambda) w_t{}^{i-1} \tag{6.11}$$

Here, λ is the exponential moving average hyperparameter, w_t is the weight of the teacher network, w_s is the weight of the student network, and i is the step number.

The student weights are updated thrice in a sequence, and the accumulated gradients in each step are backpropagated. In the first step, the model learned from precise annotated inputs of the source domain. The model is adapted for mixed inputs in the second step. In the third step, the model is learned for domain reconstruction from the masked domain inputs.

6.1.8 Loss functions

A combined Loss function L is used, which contains Cross Entropy (L_{seg}) loss and Structural Similarity Index Measure (L_{rec}) loss. The CE loss can be represented as

$$L_{seg} = \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log p_{ij} \tag{6.12}$$

6.1.8.1 Loss function in the first step

In the first step of training, the model is trained with source data. The CE loss is used for the calculation of L as per the following

$$L_1 = L_{seg} \tag{6.13}$$

6.1.8.2 Loss function in the second step

We used self-training based mixed training, which utilized ClassMixed input in the second step. CE loss is used for the calculation of L_2 for each iteration of the loop in the second

step of the training sequence.

$$L_2 = L_{seg} \tag{6.14}$$

6.1.8.3 Loss function in the third step

The proposed Masked Domain Dual Adaptation (MDDA) approach based training is done in the third step. The loss L_3 is calculated based on the following equations

$$L_3 = \alpha_1 L_{seg} + \alpha_2 L_{rec} + \beta L_{cons} + \gamma L_{dis} + \lambda L_{adv} \tag{6.15}$$

Here, L_{seg} is the segmentation loss calculated for MDSM outputs and ClassMix inputs, which are represented as

$$\begin{aligned} L_{seg} &= L_{seg_{S\to T}} + L_{seg_{T\to S}} \\ L_{rec} &= L_{rec_{S\to T}} + L_{rec_{T\to S}} \end{aligned} \tag{6.16}$$

6.1.9 Datasets

ISPRS Potsdam and Vaihingen datasets are used in four domain adaptation tasks. LoveDA dataset is used in one domain adaptation task. The Potsdam dataset has two types of inputs (RGB and IRRG), and the Vahingen dataset has a single input type (IRRG)). In ISPRS sets, adaptation is presented between *ISPRS Potsdam* and *ISPRS Veihingen* datasets. Total four adaptation tasks of Potsdam(RGB) to Vaihingen (IRRG) ($Potsdam_{RGB} \rightarrow Vaihingen_{IRRG}$), Potsdam(IRRG) to Vaihingen (IRRG) ($Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$), Vaihingen (IRRG) to Potsdam(RGB) ($Vaihingen_{IRRG} \rightarrow Potsdam_{RGB}$), and Vaihingen (IRRG) to Potsdam(IRRG) ($Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$) is used. Datasets in this task have images with different sizes, but for model training and performance comparison, patches are uniformly cropped to create the patched datasets as used in [102].

6.1.9.1 $Potsdam_{RGB} \rightarrow Vaihingen_{IRRG}$ ($P2V_{rgb}$) task

The dataset of this task has training sets of 2904 and a testing set of 440 images.

6.1.9.2 $Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$ $(P2V)$ task

The dataset of this task has training sets of 2904 and a testing set of 440 images.

6.1.9.3 $Vaihingen_{IRRG} \rightarrow Potsdam_{RGB}$ $(V2P_{rgb})$ task

This task dataset has training sets of 1296 and a testing set of 1694 images.

6.1.9.4 $Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$ $(P2V)$ task

The dataset of this task has training sets of 1296 and a testing set of 1694 images.

6.1.9.5 $LoveDA_{Rural} \rightarrow LoveDA_{Urban}$ task

This task dataset contains a total of 1366 train set images and 992 validation set images. It also consists of 820 urban images for testing.

6.1.10 Experiments

6.1.10.1 Hyperparameters and other settings

The batch size of 6, the initial learning rate of 0.0001, weight decay of 0.005, and beta values (0.9, .99) are used in the experiments. The random patches are cropped to size 512×512 during training. ADAM optimizer is utilized in all components of the MDAANet. A learning rate of 0.00001 is used for both discriminators.

6.1.10.2 Evaluation

The mean intersection over union ($mIoU$) and mean F1 Score (mF_1) are utilized to assess the measurable performance comparison of the different models. The subjective performance comparison is done using visual output assessment. The quantitative parameters are calculated for each class using the following formula

$$\begin{aligned} IoU &= \frac{tp}{tp + fp + fn} \\ F_1 &= \frac{tp}{tp + \frac{fp+fn}{2}} \end{aligned} \tag{6.17}$$

Here tp, fp, and fn are true positive, false positive, and false negative values. Average values ($mIoU$, mF_1) are calculated using each class value.

6.1.11 Assessment

The numerical and visual comparative analysis is performed with recently published works to demonstrate the effectiveness of the proposed approach.

6.1.11.1 Numerical Assessment

For numerical assessment, results produced by recent works are compared with MDAANet on quantitative parameters of $mIoU$ and mF_1 for five domain adaptation tasks.

6.1.11.2 $P2V_{rgb}$ task

Table 6.1 presented the quantitative results for $Potsdam_{IRG} \rightarrow Vaihingen_{IRRG}$ task. MDAANet achieved the best $mIoU$, mF_1 values of 64.81% and 76.8%. ST-DASegNet [102] achieved the second best values for $mIoU$ and mF_1. MDAANet attained the improvement of $\sim$ 8 % and $\sim$ 9 % respectively for $mIoU$ and mF_1 values. MDAANet achieved the best results of IoU and F_1 for four classes (Building, Low Vegetation, Tree, Car) and second best results for one class (Impervious surface). For three classes (Low Vegetation, Tree, Car), the quantitative improvements are significant in the range of $\sim$ 5-13 %, and for the Building class, MDAANet achieved an improvement of $\sim$ 1-2 %. For the Clutter class, MDAANet performs relatively worse than CCDA+LGFA [130] and Bai et al. [286].

6.1.11.3 $V2P_{rgb}$ task

Table 6.2 shown results for $Vaihingen_{IRRG} \rightarrow Potsdam_{RGB}$. MDAANet attained the highest values of 61.42% and 72.73% values for $mIoU$ and mF_1. ST-DASegNet [102] got the previous best results of 59.65%, 69.33% for mIU and mF_1 values. MDAANet improved the values of $mIoU$ and mF_1 by $\sim$ 2-3% from the earlier results. MDAANet achieved the best values of IoU and F_1 for four classes (Impervious Surface, Building, Low Vegetation, Car)

and the second best value for the Tree class. It attained improvement in the range of $\sim$ 2-5 % for IoU and F_1 vales on four classes (Impervious Surface, Buiding, Low Vegetation, Car). For the Clutter class, the performance of MDAANet is poor compared to DualGAN [287], and CCDA+LGFA [130].

6.1.11.4 $P2V$ task

Table 6.3 depicted results for $Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$ task. MDAANet achieved the best $mIoU$, mF_1 values of 74.12% and 86.24% with an improvement of 10% and 9% , respectively. PFST achieves the previous best results [102] with the value of 64.64 for $mIoU$. MDAANet got the best F_1 values for all classes and the highest IoU for four classes (Impervious Surface, Low Vegetation, Tree, Clutter). For the Building class, it attained the third best result for IoU value, which PFST precedes [120] and S&GDA [107]. It also achieved the third-best IoU result for the Car class, in which PFST precedes [120] and S&GDA [107]. MDAANet attained performance improvement in the $\sim$ 1-13 % range for F_1, with the highest improvement achieved for the Low Vegetation class.

6.1.11.5 $V2P$ task

In Table 6.4, results for $Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$ task is presented. MDAANet got 64.96%, 74.76% values for $mIoU$ and mF_1 . S&GDA achieves the second-best value for $mIoU$ [107]. MDAANet achieved the performance improvement of 2 % and 6% for $mIoU$ and mF_1 values. It achieved the highest IoU value for four classes (Impervious Surface, Building, Low Vegetation, Car) and second highest value of IoU for Tree class with improvement in the range of $\sim$ .6% - 12% with the highest improvement for Car class. For F_1 value, it attained the best results on five classes (Impervious Surface, Building, Low Vegetation, Tree, Car) in the range of $\sim$ 2% -7% with the highest improvement for Low Vegetation. It produced an inferior result for the Clutter class compared to CCDA+LGFA [130] and DualGAN [287].

6.1.11.6 *R2U* task

In Table 6.5, results for $LoveDA_{RURAL} \rightarrow LoveDA_{URBAN}$ task is shown. MDAANet attained the highest $mIoU$ value of 51.63 %, which is followed by ST-DASegNet [102] with a value of 50.28 %. MDAANet improved the $mIoU$ values by 1.4% with the highest improvement in the result of the Road class. MDAANet achieved the highest IoU for three classes (Building, Road, Water) and second highest results for three results (Background, Forest, Agriculture). JDAF achieves the highest result for the Barren class [288], which is preceded by DCA [104].

6.1.11.7 Visual Assessment

Visual comparisons of outputs of different methods are shown in Fig. 6.6, Fig. 6.8, Fig. 6.9, Fig. 6.7 and Fig. 6.10. The color scheme used for segmentation results is shown in Fig. 6.12.

6.1.11.8 *P2V*

The visual comparisons for $Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$ is shown in 6.7. The outputs of MDAANet are much better compared to other methods.

6.1.11.9 *V2P*

The comparative visual results for $Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$ is presented in 6.9. The best qualitative outputs, comparable to the ground truth, are produced by MDAANet.

6.1.11.10 $P2V_{rgb}$

In Fig. 6.6, visual results for $Potsdam_{RGB} \rightarrow Vaihingen_{IRRG}$ is shown. MDAANet can able to segment small objects that are missed by other approaches.

Dataset	Im Surface		Building		Low Vegetation		Tree		Car		Clutter		Overall	
	Iou	F_1	Iou	F_1	Iou	F_1	Iou	F_1	Iou	F_1	Iou	F_1	mIou	mF_1
AdaptSegNet [289]	2.99	5.81	51.26	67.77	10.25	18.54	51.51	68.02	12.75	22.61	60.72	75.55	31.58	43.05
ProDA [290]	2.39	5.09	49.04	66.11	31.56	48.16	49.11	65.86	32.44	49.06	68.94	81.89	38.91	52.70
DualGAN [287]	3.94	13.88	49.16	61.33	40.31	57.88	55.82	70.66	27.85	42.17	65.44	83.00	39.93	54.82
Bai et al. [286]	10.80	19.40	62.40	76.90	38.90	56.00	53.90	70.00	35.10	51.90	74.80	85.60	46.00	60.00
ResidualGAN	55.54	71.36	78.97	88.23	29.15	44.97	57.79	73.21	48.49	65.19	9.76	16.08	46.62	59.84
CCDA+LGFA [130]	12.38	21.55	64.47	77.76	43.43	60.05	52.83	69.62	38.37	55.94	76.87	86.95	48.06	61.98
GLA+CLA [106]	73.80	84.92	83.76	91.16	43.27	60.40	44.41	61.50	43.24	60.38	12.61	22.39	50.18	63.46
CIA [119]	62.63	77.02	79.71	88.71	33.31	49.97	63.43	77.62	52.28	68.66	13.50	23.78	50.81	64.29
JDAF [288]	64.33	78.29	75.53	86.06	42.16	59.31	51.99	68.41	45.87	62.90	32.71	49.30	52.10	67.38
ST-DASegNet [102]	68.36	81.28	84.09	91.33	34.69	47.08	64.65	78.31	43.15	60.28	36.03	50.64	55.16	68.15
Proposed MDAANet	72.02	84.19	89.57	94.53	58.15	71.31	73.1	80.33	56.77	77.61	39.23	52.81	64.81	76.8

Table 6.1: *Numerical results of domain adaptive semantic segmentation of $Potsdam_{RGB} \rightarrow Vaihingen_{IRRG}$ task. The red color represents the highest value and the second highest value is marked with blue color.*

Dataset	Im Surface		Building		Low Vegetation		Tree		Car		Clutter		Overall	
	Iou	F_1	Iou	F_1	Iou	F_1	Iou	F_1	Iou	F_1	Iou	F_1	mIou	mF_1
AdaptSegNet [289]	8.36	15.33	49.55	64.64	40.95	58.11	22.59	36.79	34.43	61.50	48.01	63.41	33.98	49.96
ProDA [290]	10.63	19.21	44.70	61.72	46.78	63.74	31.59	48.02	40.55	57.71	56.85	72.49	38.51	53.82
DualGAN [287]	11.48	20.56	51.01	67.53	48.49	65.31	34.98	51.82	36.50	53.48	53.37	69.59	39.30	54.71
CCDA+LGFA [130]	12.31	24.59	64.39	78.59	59.35	75.08	37.55	54.60	47.17	63.27	66.44	79.84	47.87	62.66
GLA+CLA [106]	66.11	79.75	68.63	81.32	35.47	51.85	28.64	43.51	65.45	80.17	10.84	17.49	45.86	59.74
JDAF [288]	60.05	75.04	71.42	83.33	27.79	43.39	38.74	55.84	58.64	73.93	18.09	30.63	45.79	60.38
CIA [119]	53.39	69.61	70.48	82.68	43.96	61.07	44.90	61.97	63.36	77.57	9.20	16.86	47.55	61.63
ST-DASegNet [102]	76.45	86.65	83.81	91.19	61.04	75.80	62.89	77.22	73.54	84.76	0.18	0.35	59.65	69.33
Proposed MDAANet	77.64	87.49	86.69	93.47	63.13	78.28	57.76	73.38	75.61	89.63	7.71	14.09	61.42	72.73

Table 6.2: *Numerical results of domain adaptive semantic segmentation of $Vaihingen_{IRRG} \rightarrow Potsdam_{RGB}$ task. The red color represents the highest value and the second highest value is marked with blue color.*

6.1.11.11 $V2P_{rgb}$

The visual results for $Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$ is presented in Fig . 6.9. MDAANet achieved the best visual results compared to other methods.

6.1.11.12 $R2U$

The visual results for $LoveDA_{Rural} \rightarrow LoveDA_{Urban}$ is presented in Fig. 6.10. MDAANet can detect pixels with agriculture that are missed or misclassified by other methods. MDAANet visual results are also better for all other classes than other methods.

6.1.12 Ablation Study

6.1.12.1 Network components based study

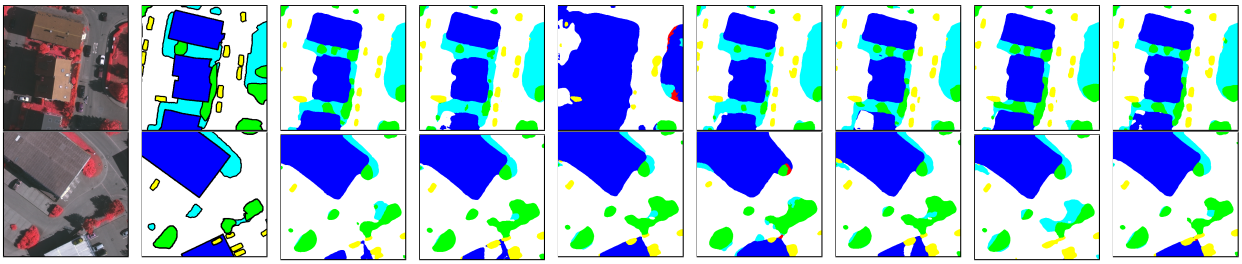
Table 6.6 depicted the ablation study of several network components. It showcased the importance of several modules used in the proposed MDAANet and their contribution to quantitative performance. All five domain adaptive semantic segmentation tasks have been

Dataset	Im Surface		Building		Low Vegetation		Tree		Car		Clutter		Overall	
	Iou	F_1	*Iou*	F_1	*Iou*	F_1	*Iou*	F_1	*Iou*	F_1	*Iou*	F_1	mIou	mF_1
AdaptSegNet [16]	4.60	8.76	54.39	70.39	6.40	11.99	52.65	68.96	28.98	44.91	63.14	77.40	35.02	47.05
FSDAN [59]	10.00	-	57.40	-	37.00	-	58.40	-	41.70	-	57.80	-	43.70	-
ProDA [290]	3.99	8.21	62.51	76.85	39.20	56.52	56.26	72.09	34.49	51.65	71.61	82.95	44.68	58.05
DualGAN [287]	29.66	45.65	49.41	66.13	34.34	51.09	57.66	73.14	38.87	55.97	62.30	76.77	45.38	61.43
Bai et al. [286]	19.60	32.80	65.00	78.80	39.60	56.70	54.80	70.80	36.20	53.20	76.00	86.40	48.50	63.10
CCDA+LGFA [130]	20.71	31.34	67.74	80.13	44.90	61.94	55.03	71.90	47.02	64.16	76.75	86.65	52.03	66.02
DRDG [291]	68.74	81.45	77.67	87.41	44.22	61.22	62.91	77.21	57.54	72.97	8.49	14.23	53.26	65.75
GLA+CLA [106]	76.58	86.73	82.74	90.56	49.97	66.64	55.22	71.15	35.44	52.33	21.85	35.87	53.63	67.21
CIA [119]	63.28	77.51	75.13	85.80	48.03	64.90	64.11	78.13	52.91	69.21	27.80	43.51	55.21	69.84
JDAF [288]	68.76	81.49	75.13	85.80	48.03	64.90	64.11	78.13	52.91	69.21	38.65	55.75	55.52	70.38
ResidualGAN [125]	72.29	83.89	80.57	89.23	49.69	66.29	63.81	77.88	57.01	72.51	11.64	18.42	55.83	68.04
S&GDA [107]	78.01	-	87.42	-	51.10	-	70.20	-	58.75	-	33.31	-	63.13	-
ST-DASegNet [102]	74.43	85.36	85.23	92.03	48.57	65.37	67.36	80.49	43.38	60.49	67.03	80.28	64.33	77.34
PFST [120]	78.85	-	87.85	-	57.30	-	62.99	-	62.11	-	38.72	-	64.64	-
Proposed MDAANet	80.9	90.39	86.89	93.17	65.66	79.56	75.54	86.74	58.55	78.12	77.17	89.43	74.12	86.24

Table 6.3: *Numerical results of domain adaptive semantic segmentation of* $Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$ *task. The red color represents the highest value and the second highest value is marked with blue color.*

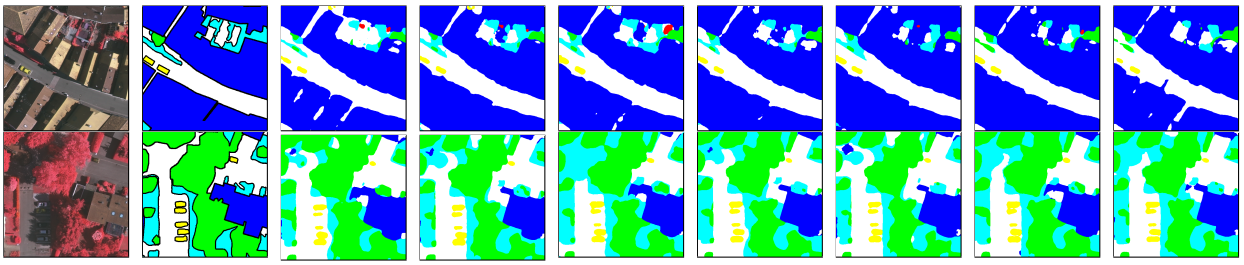
Dataset	Im Surface		Building		Low Vegetation		Tree		Car		Clutter		Overall	
	mIou	F_1	mIou	F_1	mIou	F_1	mIou	F_1	mIou	F_1	mIou	F_1	mIou	mF_1
AdaptSegNet [289]	6.11	11.50	37.66	59.55	42.31	55.95	30.71	45.51	15.10	25.81	54.25	70.31	31.02	44.75
ProDA [290]	11.13	20.51	44.77	62.03	41.21	59.27	30.56	46.91	35.84	52.75	46.37	63.06	34.98	50.76
DualGAN [287]	13.56	23.84	45.96	62.97	39.71	56.84	25.80	40.97	41.73	58.87	59.01	74.22	37.63	52.95
GLA+CLA [106]	73.43	84.55	76.32	87.43	47.69	63.45	32.68	47.36	63.86	77.85	11.65	19.47	50.94	63.31
JDAF [288]	67.70	80.74	76.36	86.59	51.19	67.72	36.21	53.17	63.22	77.47	13.10	23.17	51.30	64.81
CIA [119]	62.74	77.11	72.31	83.93	54.40	70.47	47.74	64.63	65.35	79.04	10.87	19.61	52.23	65.80
PFST [120]	71.77	-	81.59	-	57.79	-	50.44	-	66.84	-	13.27	-	56.95	-
CCDA+LGFA [130]	13.27	23.43	57.65	73.14	56.99	72.27	35.87	52.80	29.77	45.88	65.44	79.11	43.17	57.77
ST-DASegNet [102]	69.83	83.12	83.46	90.67	50.76	67.64	57.41	73.47	75.99	87.89	3.70	7.38	56.86	68.37
S&GDA [107]	76.52	-	84.28	-	62.03	-	66.03	-	72.51	-	11.34	-	62.12	-
Proposed MDAANet	81.62	90.13	84.99	92.69	64.56	79.95	65.77	79.61	84.17	92.61	8.64	13.6	64.96	74.76

Table 6.4: *Numerical results of domain adaptive semantic segmentation of* $Vaihingen_{IRGB} \rightarrow Potsdam_{IRRG}$ *task. The red color represents the highest value and the second highest value is marked with blue color.*



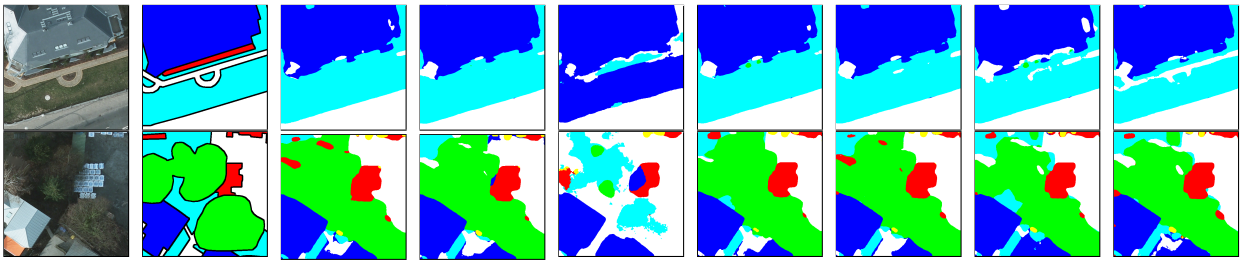
(a) *Input* **(b)** *Label* **(c)** *AdaptSeg* **(d)** *DualGAN* **(e)** *CIA* **(f)** *PFST* **(g)** *ProDA* **(h)** *ST-DA* **(i)** *Our*

Figure 6.6: *Visual comparison of output produced by several models on* 512×512 *pixels patches for* $Potsdam_{RGB} \rightarrow Vaihingen_{IRRG}$ *task.*



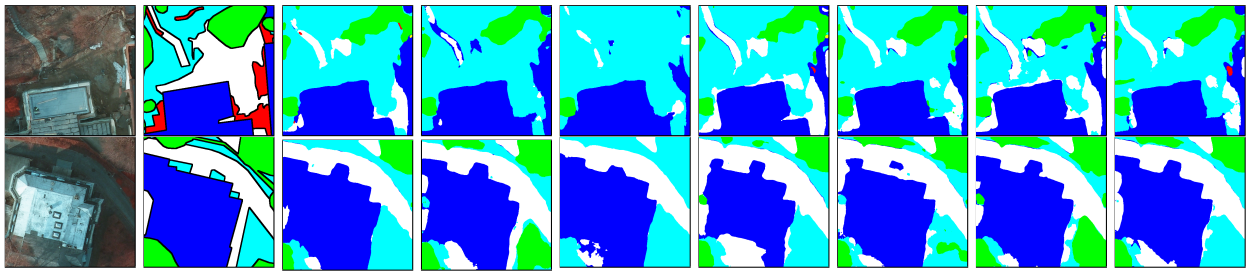
(a) *Input* (b) *Label* (c) *AdaptSeg* (d) *DualGAN* (e) *CIA* (f) *PFST* (g) *ProDA* (h) *ST-DA* (i) *Our*

Figure 6.7: *Visual comparison of output produced by several models on* 512×512 *pixels patches for* $Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$ *task.*



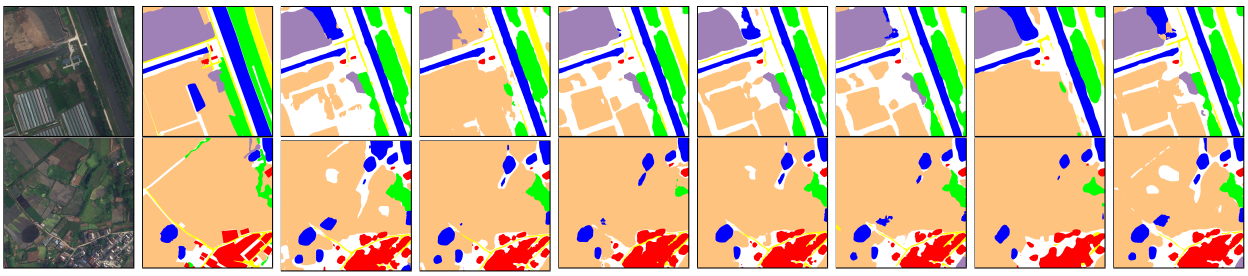
(a) *Input* (b) *Label* (c) *AdaptSeg* (d) *DualGAN* (e) *CIA* (f) *PFST* (g) *ProDA* (h) *ST-DA* (i) *Our*

Figure 6.8: *Visual comparison of output produced by several models on* 512×512 *pixels patches for* $Vaihingen_{IRRG} \rightarrow Potsdam_{RGB}$ *task.*



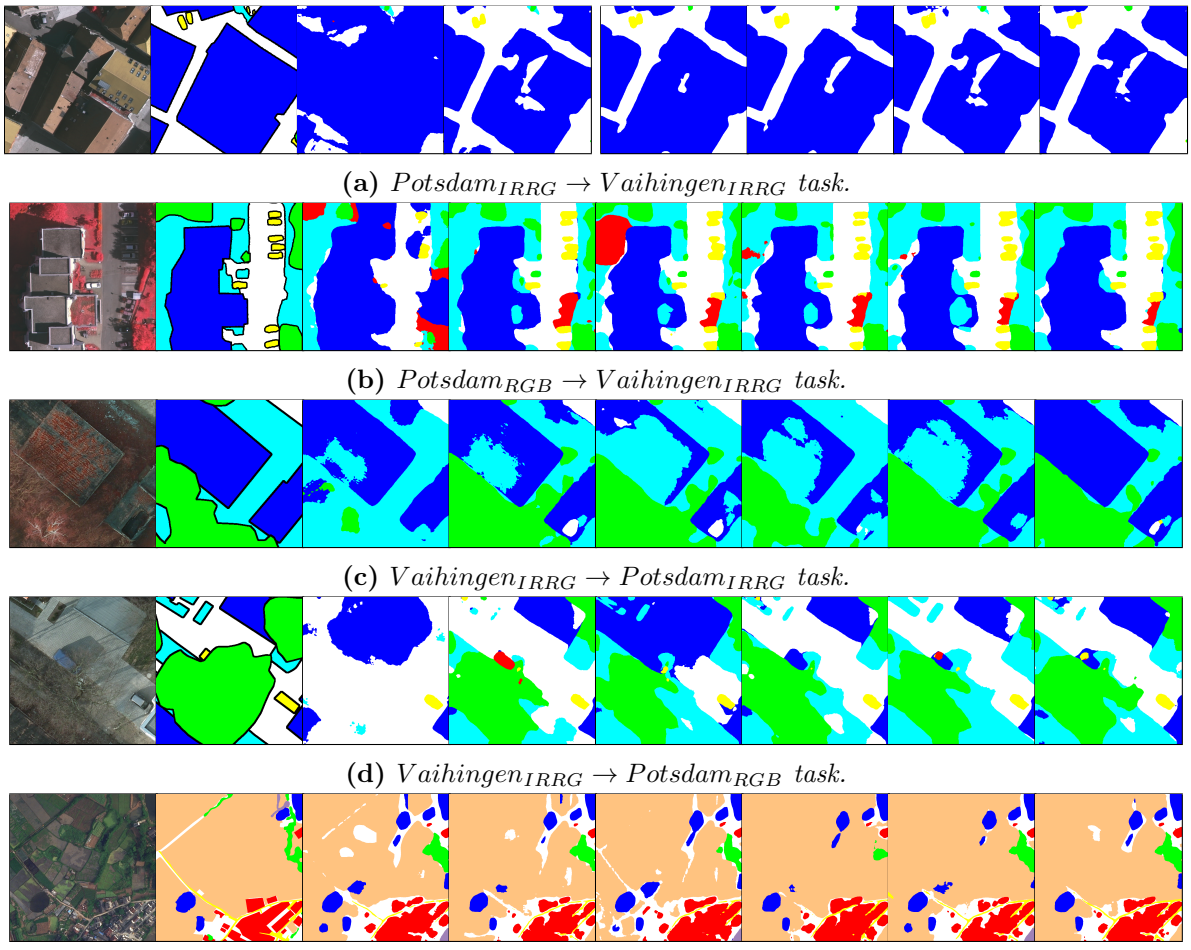
(a) *Input* (b) *Label* (c) *AdaptSeg* (d) *DualGAN* (e) *CIA* (f) *PFST* (g) *ProDA* (h) *ST-DA* (i) *Our*

Figure 6.9: *Visual comparison of output produced by several models on* 512×512 *pixels patches for* $Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$ *task.*



(a) *Input* (b) *Label* (c) *AdaptSeg* (d) *DualGAN* (e) *CIA* (f) *PFST* (g) *ProDA* (h) *ST-DA* (i) *our*

Figure 6.10: *Visual comparison of output produced by several models on* 512×512 *pixels patches for* $LoveDA_{Rural} \rightarrow LoveDA_{Urban}$ *task.*



(a) $Potsdam_{IRRG} \rightarrow Vaihingen_{IRRG}$ *task.*

(b) $Potsdam_{RGB} \rightarrow Vaihingen_{IRRG}$ *task.*

(c) $Vaihingen_{IRRG} \rightarrow Potsdam_{IRRG}$ *task.*

(d) $Vaihingen_{IRRG} \rightarrow Potsdam_{RGB}$ *task.*

(e) $LoveDA_{RURAL} \rightarrow LoveDA_{URBAN}$ *task.*

Figure 6.11: *Visual comparison of ablation study of network components on* 512×512 *pixels patches for five domain adaptation tasks. Starting from the leftmost column, images are shown for Input, Label, B+ST, B+ST+MDA, B+ST+MDAA, B+ST+MDAA+CE, B+ST+MDAA+CE+JAA, B+ST+MDAA+CE+JAA+FDA.*

	Background	Building	Road	Water	Barren	Forest	Agriculture	Mean
AdaptSeg [289]	42.35	23.73	15.61	81.95	13.62	28.70	22.05	32.68
DCA [104]	45.82	49.60	51.65	80.88	16.70	42.93	36.92	46.36
JDAF [288]	49.75	42.62	39.55	70.79	36.12	39.71	51.29	47.12
FPN [105]	50.20	49.50	43.86	86.90	15.00	42.67	42.51	47.23
SMCDA [292]	51.79	50.03	52.16	79.63	15.55	43.72	38.40	47.66
HIUDA [101]	49.30	55	55.40	86	17.10	41.20	36.90	48.70
ST-DASegNet [102]	51.01	54.23	60.52	87.31	15.18	47.43	36.26	50.28
Ours	51.44	55.01	63.10	88.32	15.13	46.01	42.42	51.63

Table 6.5: *Numerical results of domain adaptive semantic segmentation of $LoveDA_{Rural} \rightarrow LoveDA_{Urban}$ (R2U) task. The red color represents the highest value and the second highest value is marked with blue color.*

	$P2V_{rgb}$		$P2V$		$V2P$		$V2P_{rgb}$		$R2U$
	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$
B	39.36	46.41	57.63	73.01	54.26	63.97	38.39	48.97	41.77
B+ST	60.61	72.89	68.6	81.12	59.05	69.79	55.83	67.53	43.32
B+ST+MDA	60.28	73.16	71.71	83.3	61.85	70.9	56.59	67.86	44.54
B+ST+MDDA	61.3	73.78	72.35	84.02	62.32	71.37	57.0	68.26	46.99
B+ST+MDDA+CE	64.15	75.55	73.68	84.78	62.58	72.1	57.35	68.76	48.76
B+ST+MDDA+CE+JAA	65.46	77.42	74.06	85.36	63.84	73.33	59.91	70.74	50.79
B+ST+MDDA+CE+JAA+FDA	65.93	78.47	74.12	86.24	64.96	74.76	61.89	72.98	51.63

Table 6.6: *Numerical results of domain adaptive semantic segmentation for five domain adaptation tasks.*

selected for this study. The initial Bases model (B) is selected, trained, and evaluated on the target domain. Later, Self Training for the target domain is added to the training process (B+ST). Mix Domain Adaptation (B+ST+MDA) further improves the performance of the models. Mix domain dual adaptation (B+ST+MDDA) enhanced the results from the previous values. Consistency Enforcment (B+ST+MDDA+CE) further improved the results by 0.5-3 %. Joint Adversarial Alignment (B+ST+MDDA+CE+JAA) achieved performance improvement in the range of 0.5-2 %. Later, the FDA attained the final values.

	$P2V_{rgb}$		$P2V$		$V2P$		$V2P_{rgb}$		$R2U$
	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$
NA	60.61	72.89	68.6	81.12	59.05	69.79	55.83	67.53	43.32
DD	61.21	73.13	69.93	81.14	59.26	70.09	57.08	68.39	47.99
MDD	62.31	74.19	70.34	81.34	60.81	71.72	58.99	69.19	49.34
MDD+JAA	65.46	77.42	74.06	85.36	63.84	73.33	59.91	70.74	50.79

Table 6.7: *Ablation results for adversarial components. Here, NA is the network with no adversarial component, DD is the network with a domain discriminator, MDD is the network with a masked domain discriminator, and JAA represents the joint adversarial alignment.*

	$P2V_{rgb}$		$P2V$		$V2P$		$V2P_{rgb}$		$R2U$
	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$
NR	60.61	72.89	68.6	81.12	59.05	69.79	55.83	67.53	43.32
CR	60.97	72.99	69.87	82.18	59.83	70.12	55.92	67.66	43.62
SR	61.05	73.03	70.43	82.72	60.45	70.42	56.07	67.74	43.98
MDR	61.21	73.16	71.71	83.3	61.85	70.9	56.59	67.86	44.52
MDDR	61.3	73.78	72.35	84.02	62.32	71.37	57.0	68.26	46.99

Table 6.8: *Ablation results for the combination of Reconstruction Modules. Here, NR is the network with no reconstruction module, CR represents a network with a common reconstruction module, SR is a network with separate reconstruction modules, DR is a network with separate domain reconstruction, MDR is the network with separate mask domain reconstruction, and MDDR is MDR with dual reconstruction.*

	$P2V_{rgb}$		$P2V$		$V2P$		$V2P_{rgb}$		$R2U$
	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$
FS	39.36	46.41	57.63	73.01	54.26	63.97	38.89	48.97	41.77
FS+SS	60.61	72.89	68.6	81.12	59.05	69.79	55.83	67.53	43.32
FS+TS	60.93	72.96	68.51	80.93	59.35	69.91	55.96	67.91	45.02
FS+SS+TS	65.93	78.47	74.12	86.24	64.96	74.76	61.89	72.98	51.63

Table 6.9: *Ablation results for the training stages. Here, FS is the first stage of training, SS represents the second stage of training, and TS is the third stage of training.*

α_1	α_2	β	γ	λ	$P2V_{rgb}$		$P2V$		$V2P$		$V2P_{rgb}$		$R2U$
					$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$	mF_1	$mIou$
1	0	0	0	0	60.61	72.89	68.6	81.12	59.05	69.79	55.83	67.53	43.32
1	0.75	0	0	0	61.3	73.78	72.35	84.02	62.32	72.37	57.0	68.26	46.99
0.25	0.75	0	0	0	61.1	73.32	71.53	82.13	59.82	70.11	56.08	67.92	45.72
1	0.5	0	0	0	62.12	75.54	72.91	84.31	62.86	72.74	58.78	69.97	47.44
0.5	0.5	0	0	0	60.73	72.92	68.94	81.18	59.86	69.95	55.99	67.43	43.44
1	0.25	0	0	0	62.73	75.69	73.37	85.05	63.29	72.89	60.28	69.88	47.79
0.75	0.25	0	0	0	60.73	72.92	68.94	81.18	59.86	69.95	55.99	67.43	43.44
1	0.25	0.1	0	0	64.15	75.55	73.68	85.11	63.58	73.1	60.87	70.31	48.76
1	0.25	0.2	0	0	63.95	74.21	73.31	84.95	63.55	73	60.46	69.92	48.76
1	0.25	0.1	0	0.1	65.46	77.42	74.06	85.36	63.84	73.33	59.91	70.74	50.79
1	0.25	0.1	0	0.2	65.13	77.09	74.01	85.18	63.54	73.11	59.76	70.18	50.79
1	0.25	0.1	0.1	0.1	65.93	78.47	74.12	86.24	64.96	74.76	61.89	72.98	51.63
1	0.25	0.1	0.2	0.1	64.81	76.8	74.12	86.24	64.96	74.76	61.42	72.73	50.22

Table 6.10: *Ablation results for optimal value of loss parameter α_1, α_2, β, γ, and λ.*

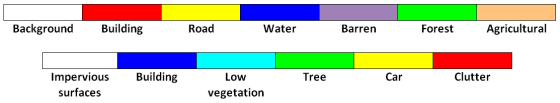


Figure 6.12: *Color code used in experiments. The top row is for the LoveDA dataset, and the bottom row is for the ISPRS dataset.*

6.1.12.2 Adverserial module based study

In Table 6.7, an ablation study for various adversarial components of MDAANet is presented. The results are calculated for all five domain adaptation tasks. The initial results are generated without the use of any adversarial component. The addition of a domain discriminator (between source and target reconstructed output) resulted in the improvement of $\sim$ .2 - 1.6 for $mIoU$ and of $\sim$ .02 - .2 % for mF_1 values. Masked domain discriminator further improved the overall performance of adaptation tasks. Finally, the use of joint adversarial alignment on masked domain discriminator produced the improvement of $\sim$ 1 - 3 % for $mIoU$ and $\sim$ 1 - 4 % for mF_1 values.

6.1.12.3 Reconstruction module based study

The ablation study for several components of reconstruction modules is presented in Table 6.8. The initial results are produced without using any reconstruction component. The use of a common reconstruction module for source and target resulted in the improvement of $\sim$ .09 -1.5 % in $mIoU$ and $\sim$ .1-1 % in mF_1 values. The utilization of a separate reconstruction module further improved the results and produced better improvement compared to the common reconstruction module. Masked domain reconstruction attained quantitative enhancement of $\sim$ .05 -.3 % for $mIoU$ and $\sim$ .12 - 5 % for mF_1. Masked domain dual reconstruction improved the results by .07 - 1.5 % for $mIoU$ and $\sim$ .6 - 1.4 % for mF_1.

6.1.12.4 Training stages based study

Table 6.9 depicted the ablation results for the optimum number of training stages. The study started with only the first stage of training and produced results for all five domain adaptation tasks. The addition of second-stage training improved results by $\sim$ 5-20 % for

$mIoU$ and $\sim$ 6-25 % for mF_1 values. Combining the first stage with third stage improved results further for three tasks ($P_{RGB} \rightarrow V_{IRRG}$, $V_{IRRG} \rightarrow P_{IRRG}$, $V_{IRRG} \rightarrow P_{RGB}$), and results marginally deteriorated for ($P_{IRRG} \rightarrow V_{IRRG}$) from the previous step. Combined training with all three stages resulted in performance improvement of $\sim$ 5-6 % for $mIoU$ and mF_1 from the previous steps. It is concluded that the best quantitative results are produced when all three stages are used for training.

6.1.13 Loss weight parameters study

Three loss weight parameters of α_1, α_2, β, γ and λ are used in the proposed study. The results of the ablation study are presented in Table 6.10. Based on the analysis of the results of this section, the values of $\alpha_1, \alpha_2, \beta$, γ and λ are chosen. We experimentally observed that the best results are achieved using values of 1.0, 0.25, 0.1, 0.1, and 0.1

6.1.13.1 Ablation Study visualization

The visual comparison for outputs of the ablation study of network components for five domain adaptation tasks is presented in Fig. 6.11. The figures represented the importance of each network component in the final results. The addition of each component improves the qualitative results from the previous output. The initial results of the base network with self-training are kept as the baseline results. The visual results for the inclusion of Masked Domain Adaptation (MDA), Masked Domain Dual Adaptation (MDDA), Consistency Enforcement (CE), Joint Adversarial Alignment (JAA), and Feature Dissimilarity based Alignment (FDA) in the baseline are depicted. It reveals that these modules improve the baseline results.

6.2 RADA

OT in nighttime conditions is important, and several approaches [209,210,213,214] used LLE based nighttime detection and tracking. It has been observed in the literature that visual

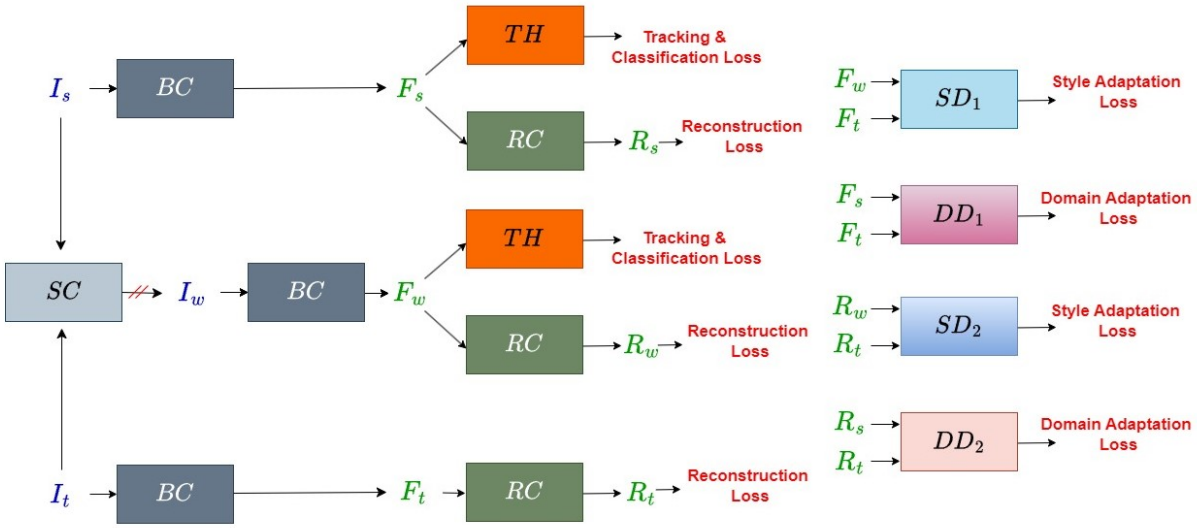


Figure 6.13: *Illustration of the proposed RADA framework. Here, I_s, I_t, and I_w are search and template image pairs of source, target, and styled source domains. F_s, F_t, and F_w are extracted source, target, and styled source features using the Backbone Component (BC).*

OT at night time is a quite challenging task due to the scarcity of paired (labeled) images to train the supervised models. Domain adaptation is a technique that extends transfer learning by minimizing the differences between the domain-specific features across different domains. These approaches have gained popularity in semantic segmentation [216, 220], object detection [210, 215] and, recently, in OT [212, 217, 218]. Several strategies involving domain adaptation have been developed, including addressing domain shifts at both image and instance levels [215] and merging low-light enhancement and object detection models [210]. Some works [217, 218] used adversarial learning for domain adaptation in visual OT to minimize the domain discrepancy between day and night features and showed promising results in addressing the domain shift problem in OT.

We proposed static image style transfer-based RADA with adversarial learning for night-time OT. RADA attained feature and input level adaptation, which resulted in better adaptation without external model requirements for low-light image enhancement. The major contributions are as follows:

1. Reconstruction assisted adaptation is proposed for domain invariant feature extraction and to attain input and feature level adaptation.

2. Style adversarial alignment at multiple level is proposed to adapt between the styled source domain and the target domain, which do not require pseudo labels.

RADA achieved SOTA results on two benchmark dataset for domain adaptive nighttime tracking.

6.2.1 Proposed Framework

The various components of RADA are presented in Fig. 6.13. It consists of a Backbone Component (BC), a Tracking Head (TH), a Reconstruction Component (RC), and Discriminators (DD_1, DD_2, SD_1, SD_2). Adding these modules aims to improve the robustness of the tracker in adverse conditions, enhance domain adaptation, and utilize the benefits of style transfer for OT.

Template and search images are obtained from the input video sequences. BC utilized these inputs and extracts features that TH used for regression and classification when labeled inputs (source image and styled source image) are passed. The feature maps are also passed (in all cases) into the proposed RC, which learns to reconstruct the input images. Discriminators used extracted features and reconstructed outputs for alignment.

6.2.1.1 Backbone Component

The backbone component (BC) is used to extract features from the input. We used mixed attentive transformer [207] based common BC for source, target, and styled source inputs. Inputs having search image of size $3 \times 384 \times 384$ template image of size $3 \times 192 \times 192$ are passed to BC for feature extraction. The extracted features have sizes of $1024 \times 24 \times 24$ and $1024 \times 12 \times 12$ for search and template inputs.

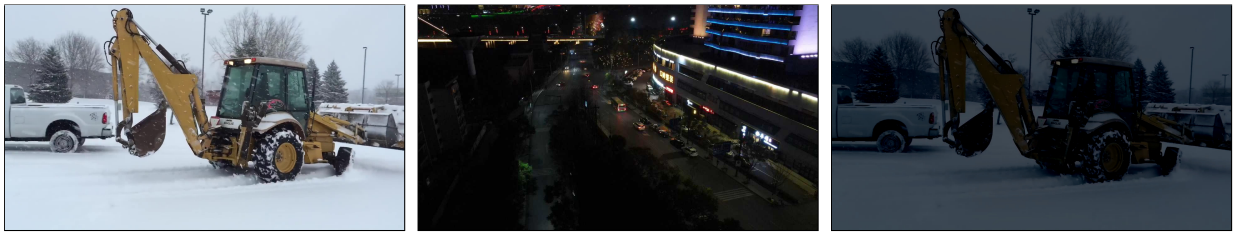
6.2.1.2 Style transfer Component

We used Wallis Style Transfer [199, 293] for static style transfer of the target domain night image to the source domain day image. It works by blending the style image over the content image as shown in 6.14. For source image I_s, target image I_t, target styled source image is

produced using following equation :

$$I_w = w[\sigma(I_t)(\frac{I_s - \mu(I_s)}{\sigma(I_s)}) + \mu(I_t)] + (1 - w)I_s \tag{6.18}$$

Here, w is the weight parameter, μ is the mean, and σ is the variance. The style source image I_w with the same annotation is used for model training and extracting domain invariant features. We used $w = 1$ in experimentation for a complete nighttime style transfer.



(a) *GOT image* **(b)** *NAT2021 image* **(c)** *Output styled image*

Figure 6.14: *Visualization of style transfer.*

6.2.1.3 Reconstruction Component

Reconstruction Component (RC) uses multitask learning which helps to learn a common representation of the domains and to attain input-level domain adaptation through reconstruction. Reconstruction Component consists of transposed convolution-based upsampling with a resolution module. It takes features extracted by BC and reconstructs the original passed inputs. RC takes input shapes of $1024 \times 32 \times 32$ and $1024 \times 16\times$ for search and template input and produces a reconstructed output of shape $3 \times 384 \times 384$ and $3 \times 192 \times 192$.

6.2.1.4 Tracker Head

Tracker Head (TH) consists of classification and regression branches for object presence identification and bounding box detection. TH takes feature extracted by BC with the shape of $1024 \times 32 \times 32$ and $1024 \times 16 \times 16$ for search and template inputs.

6.2.1.5 Domain Discriminators

Domain Discriminators (DD) distinguished source and target at the feature and reconstructed output levels. DD enables feature and input level adaptation between the source and target domain and helps in domain invariant feature representation. DD consists of two discriminators (DD1, DD2) for feature and reconstructed outputs. DD1 takes domain features extracted by BC of shape $1024 \times 32 \times 32$ and $1024 \times 16 \times 16$ for search and template image and produce domain classification output. DD2 used output generated by RC of shape $3 \times 192 \times 192$ and $3 \times 384 \times 384$ for template and search inputs to generate classification output.

6.2.1.6 Style Discriminators

Style Discriminators (SD) differentiate between styled source and target at the feature and reconstructed input levels. SD attains feature and input level adaptation and minimizes the difference between target-styled source domain inputs with target domain inputs. SD contains two style discriminators (SD1, SD2). SD1 took BC extracted features for template and search inputs having the shape of $1024 \times 32 \times 32$ and $1024 \times 16 \times 16$. SD2 utilized RC outputs of template and search inputs with shapes of $3 \times 192 \times 192$ and $3 \times 384 \times 384$ for style nonstyle classification.

6.2.2 Loss function

The total loss (L) for the network in one iteration can be calculated as follows:

$$L = Ls_{tra} + \alpha Ls_{rec} + \beta(Lw_{tra} + \alpha Lw_{rec}) + \alpha Lt_{rec} + \gamma L_{adv} \tag{6.19}$$

Here Ls_{tra} and Lw_{tra} are tracking loss for source and styled source inputs, which includes classification and regression loss. Ls_{rec}, Lw_{rec}, Lt_{rec} are reconstruction losses on the source, styled source and target inputs. L_{adv} is adverserial loss and loss weight parameters are represented by α, β, and λ.

6.2.3 Implementation details

6.2.3.1 Dataset

This work uses the GOT [294] dataset as a source domain dataset. For the target domain, NAT2021 benchmark [217] data set NAT2021-train is used without the labels. Performance evaluation is done on NAT2021-test and NAT2021_L-test datasets. NAT2021-train dataset contains 1400 videos with 276081 frames without annotations. NAT2021-test and NAT2021_L-test datasets have 180, 23 videos and 140815, 53564 frames respectively. NAT2021_L-test involves much longer videos and is used for the long-term tracking evaluation of the trackers. Long-term tracking is a typical scene in visual OT and involves many more challenging attributes.

6.2.3.2 Hyperparameters and evaluation metrics

The model is implemented in the PyTorch framework and experiments are run on the system with a V100 (32 GB) graphics card. Models are trained using the SGD with momentum optimizer, with the initial learning rate set to $\gamma = 0.0015$ and the momentum set to $\mu = 0.9$. The learning rate is scheduled logarithmically over the number of trained epochs, starting from 0.0015 to 0.00015. The discriminator is trained using the Adam optimizer with the initial decay rates set to $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The base learning rate is set at 0.005. The entire training is done for a total of 30 epochs. For evaluation, we use the One Pass Evaluation style. We use the same metrics as the baseline tracker for comparison: success rate, precision, and normalized precision.

6.2.4 Performance Comparison

6.2.4.1 Numerical Comparison

6.2.4.2 NAT2021-test

A performance comparison of our model with other recent work is presented in Table 6.11. RADA got the highest success rate and normalized precision values of 0.561 and 0.682. For

precision, RADA achieved the second-best result of 0.721, preceded by PDST [218] results of 0.76. The success plots and precision plot comparisons of the mentioned trackers are shown in 6.15. RADA raises the previous results of success rate and normalized precision by **2%**.

6.2.4.3 NAT2021L-test

Table 6.12 depicted the comparative results on a large sequence video dataset. RADA significantly outperforms the baseline. In success rate, RADA (0.532) raises the previous result [218] (0.507) by **4.9%**. In precision, RADA (0.681) raises the previous result [218] (0.649) by **4.9%**. The success plots and precision plot comparisons of the mentioned trackers are shown in 6.16.

6.2.4.4 Qualitative Comparison

We illustrate some instances from both datasets where our tracker is able to correctly identify the target object and locate it while the other tracker starts making arbitrary bounding boxes or tracking a different object altogether. For visual comparisons, baseline results of UDAT-CAR [217], UDAT-BAN [217], SiamCAR [205], and SiamBAN [203] are used. Fig. 6.17 visualized results for NAT2021-test datasets. Visual comparison for NAT2021_L-test dataset is shown in Fig. 6.18.

Table 6.11: *Results - NAT2021-test*

	Source Dataset	Success Rate	N. Precision	Precision
RADA (ours)	GOT [294]	0.561	0.682	0.721
PDST [218]	GOT [294], LaSOT [295]	0.547	0.665	0.76
MT-CAR [226]	GOT [294], VID [296]	0.507	0.592	0.72
MT-BAN [226]	GOT [294], VID [296]	0.494	0.562	0.699
UDAT-CAR [217]	GOT [294], VID [296]	0.483	0.564	0.687
UDAT-BAN [217]	GOT [294], VID [296]	0.469	0.546	0.694
SiamCAR [205]	GOT [294], VID [296]	0.453	0.542	0.663
SiamBAN [203]	GOT [294], VID [296]	0.437	0.509	0.647

Table 6.12: *Results - NAT2021L-test*

	Source Dataset	Success Rate	N. Precision	Precision
RADA (ours)	GOT [294]	0.532	0.625	0.681
PDST [218]	GOT [294], LaSOT [295]	0.507	0.599	0.649
MT-BAN [226]	GOT [294], VID [296]	0.399	0.455	0.556
MT-CAR [226]	GOT [294], VID [296]	0.39	0.442	0.543
SFDT [223]	GOT [294], VID [296]	0.401		0.524
UDAT-CAR [217]	GOT [294], VID [296]	0.376	0.413	0.506
UDAT-BAN [217]	GOT [294], VID [296]	0.352	0.406	0.496
SiamCAR [205]	GOT [294], VID [296]	0.330	0.375	0.477
SiamBAN [203]	GOT [294], VID [296]	0.316	0.366	0.464

Table 6.13: *Ablation study for NAT2021 dataset*

	Success Rate	N. Precision	Precision
BC+SC+RC1+RC2+DD+SD	0.561	0.682	0.721
BC+SC+RC1+RC2+DD	0.552	0.669	0.715
BC+SC+RC1+RC2	0.548	0.667	0.710
BC+SC+RC1	0.546	0.661	0.708
BC+SC	0.545	0.658	0.706
BC	0.532	0.625	0.681

6.2.5 Ablation Study

To demonstrate the effectiveness of the various components of RADA and to find optimal parameter values, extensive ablation comparisons are made on the NAT2021-test and NAT2021L-test datasets.

6.2.5.1 Ablation for network components

Table 6.13 and Table 6.14 presented network ablation studies for RADA. The initial results of BC are taken as a baseline for comparison. The addition of SC improved the qualitative results for both datasets. Including a reconstruction component for the source domain and styled source inputs (RC1) further improved the results. We extended the reconstruction component for (RC2) for target input, which further enhanced the results. The addition of adversarial components (style and domain discriminators) significantly improved the previous results produced by BC+SC+RC1+RC2. The visual results of this ablation study are shown in Fig. 6.19.

Table 6.14: *Ablation study for NAT2021_L dataset*

	Success Rate	N. Precision	Precision
BC+SC+RC1+RC2+DD+SD	0.532	0.625	0.681
BC+SC+RC1+RC2+DD	0.524	0.613	0.681
BC+SC+RC1+RC2	0.508	0.599	0.653
BC+SC+RC1	0.502	0.587	0.642
BC+SC	0.505	0.599	0.633
BC	0.491	0.564	0.621

Table 6.15: *Ablation study for loss hyperparameter for NAT2021 dataset*

α	β	Success Rate	N. Precision	Precision
0	0	0.532	0.625	0.681
0	0.25	0.545	0.658	0.706
0	0.5	0.541	0.652	0.691
0	0.75	0.539	0.649	0.689
0	1	0.537	0.641	0.681
0.25	0.25	0.548	0.667	0.710
0.25	0.5	0.546	0.661	0.706

6.2.5.2 Ablation for loss parameters

Table 6.15 and Table 6.16 depicted the ablation results for loss weight parameters of α and β. Initially, $\beta = 0$ results are produced with $\alpha = 0$. Experiments are done with increments of 0.25 in both values. From these experiments, we concluded that the best results are achieved using values of 0.25 and 0.25 for α and β.

6.3 Summary

We proposed MDAANet with Mask Domain Dual Adaptation approach (MDDA), Joint Adversarial Adaptation (JAA), Feature Dissimilarity based alignment (FDA), and Consistency Enforcement (CE) for effective domain adaptive semantic segmentation. MDDA allows input and output level domain alignment using masked domain inputs. JAA achieves input-output level combined domain adaptation. The feature-level adaptation is attained through common FEM. Consistency Enforcement (CE) ensures the reconstructed input-output pairs are similar. MDAANet achieved SOTA results on five domain adaptive semantic segmentation tasks. Detailed experimentation is presented for ablation studies of several components

Table 6.16: *Ablation study for loss hyperparameter for NAT_L2021 dataset*

α	β	Success Rate	N. Precision	Precision
0	0	0.491	0.564	0.621
0	0.25	0.505	0.599	0.633
0	0.5	0.501	0.587	0.638
0	0.75	0.5	0.579	0.634
0	1	0.49	0.569	0.627
0.25	0.25	0.508	0.599	0.653
0.25	0.5	0.506	0.587	0.651

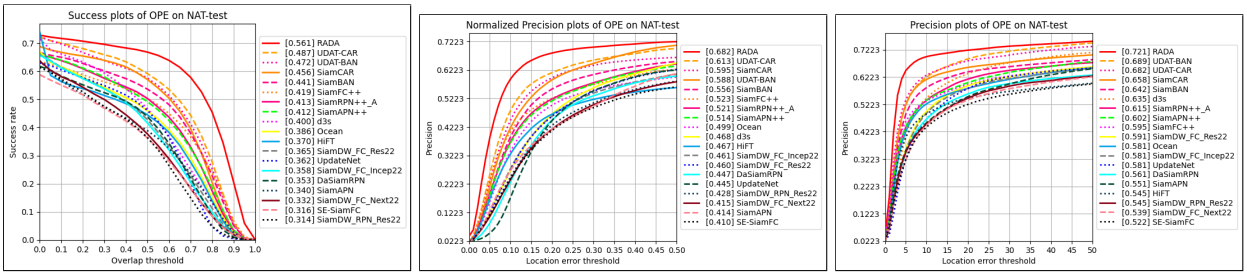


Figure 6.15: *Success, Normalized precision, and Precision plots of several trackers on NAT2021 test dataset.*

and hyperparameters of MDAANet. The performance is validated through qualitative and quantitative comparisons with recently published work which reveals that the proposed scheme outperforms SOTA schemes in most of the settings.

This work proposes reconstruction component-based domain adaptation, which results in input-level and feature-level adaptation. The style transfer component helped in learning target domain representation without generating noisy pseudo levels. Our proposed framework does not require an external nighttime image enhancement model and can produce good results without this step. Detailed qualitative and quantitative evaluations are presented for the proposed model. The results are validated on two benchmark tasks and compared with other recent works. In the next chapter, the conclusion of our contributory works and future prospective details will be presented.

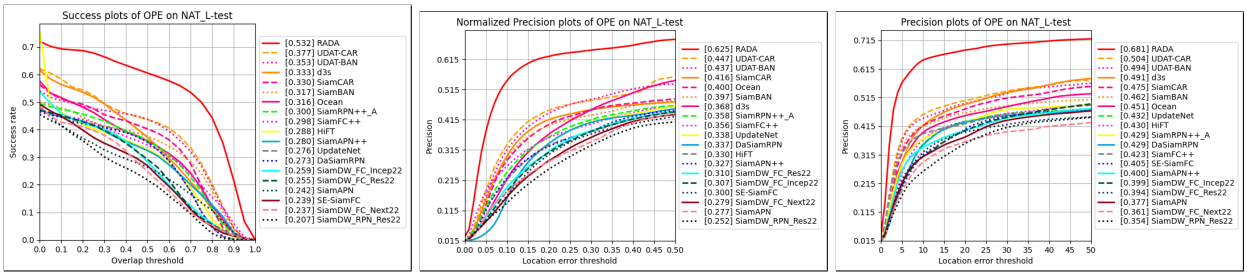


Figure 6.16: *Success, Normalized precision, and Precision plots of several trackers on NAT2021_L test dataset.*

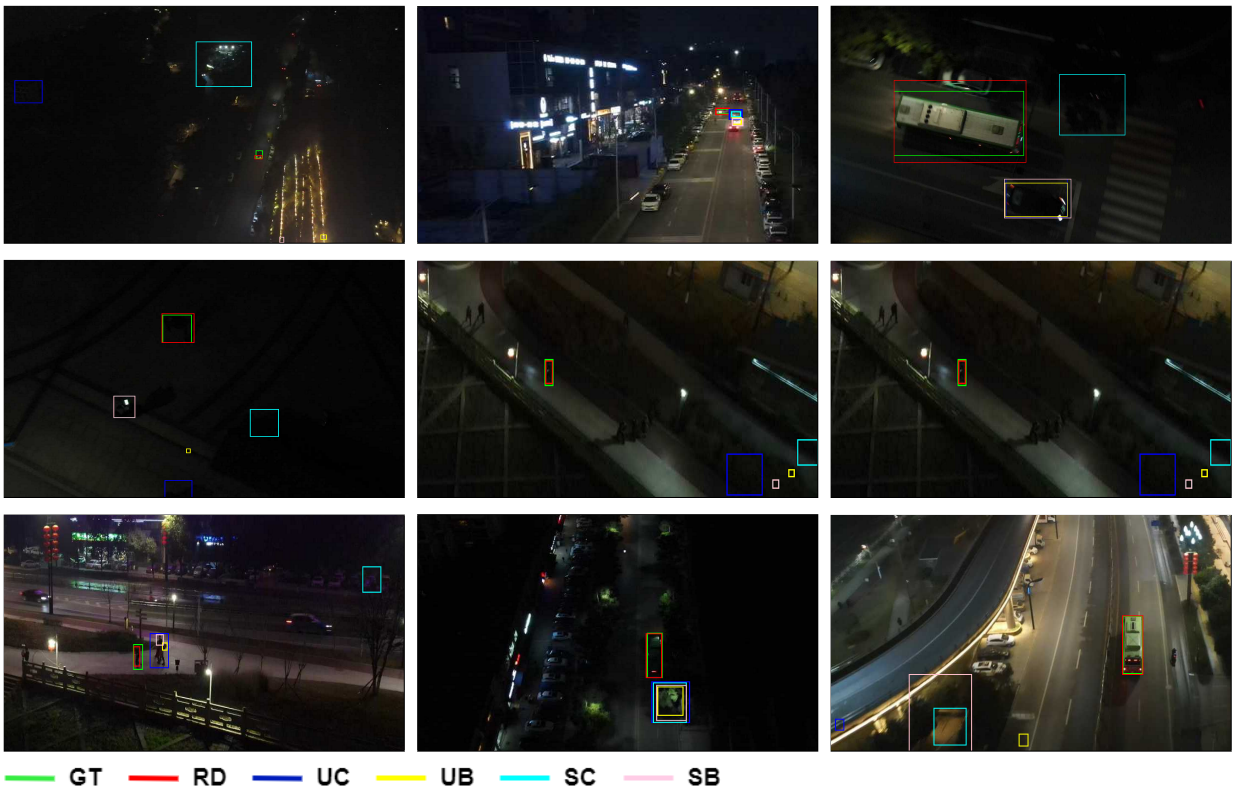


Figure 6.17: *Visual comparison of results of RADA and baselines on NAT2021-test dataset. Here, GT is ground truth, RD is RADA, UC is UDAT-CAR, UB is UDAT-BAN, SC is SiamCAR, and SB is SiamBAN.*

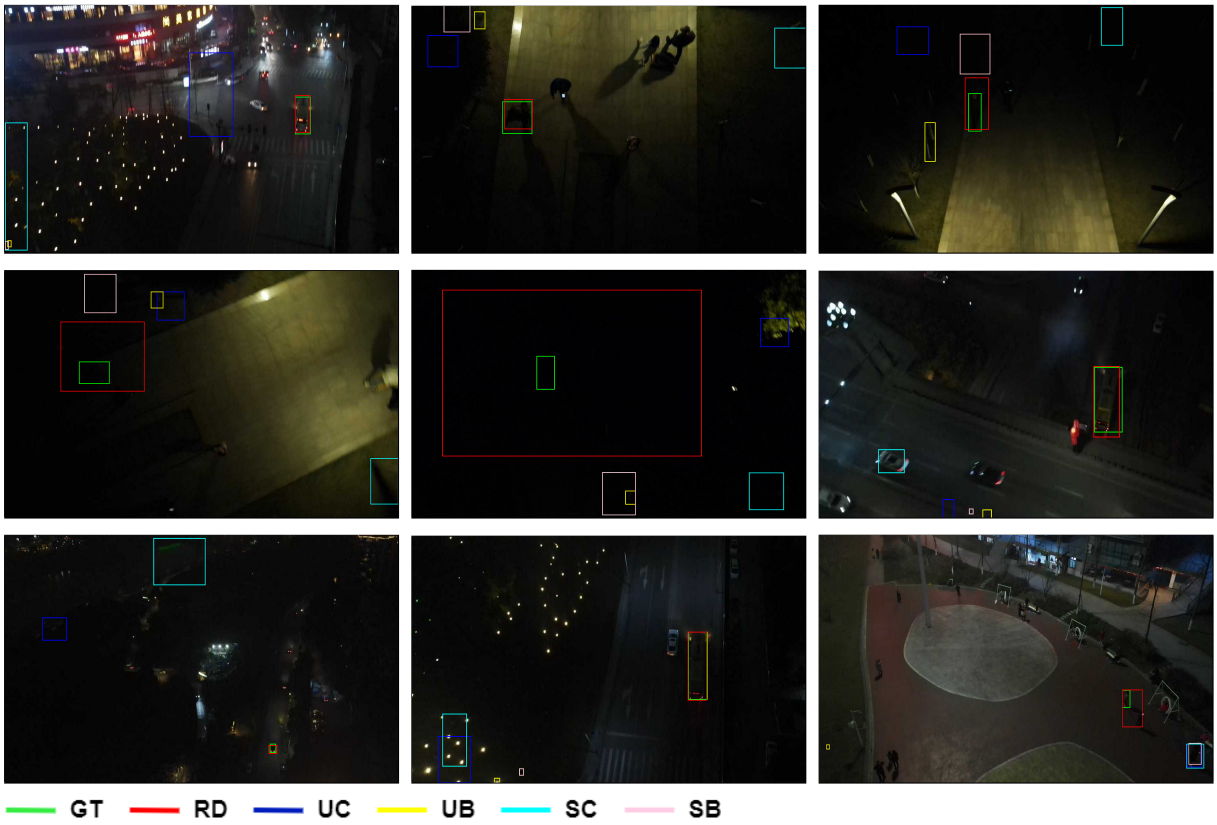
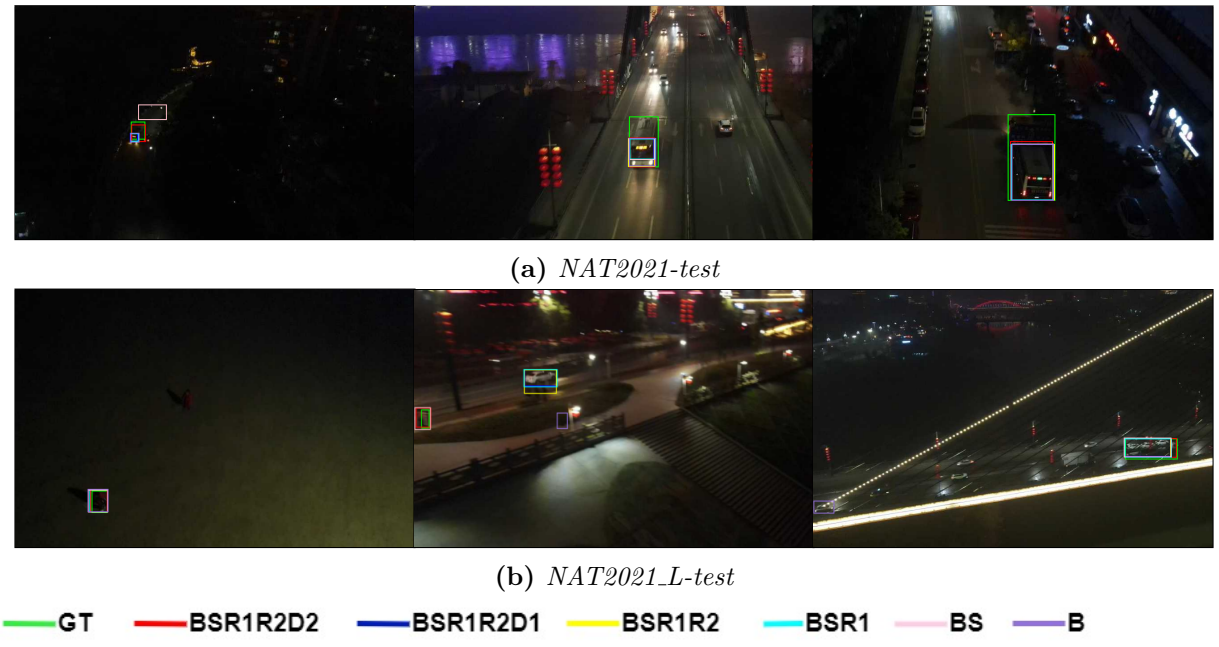


Figure 6.18: *Visual comparison of results of RADA and baselines on NAT2021_L-test dataset. Here, GT is ground truth, RD is RADA, UC is UDAT-CAR, UB is UDAT-BAN, SC is SiamCAR, and SB is SiamBAN.*



(a) *NAT2021-test*

(b) *NAT2021_L-test*

Figure 6.19: *Visual comparison of ablation study on NAT2021-test and NAT2021_L-test datasets.*

"Let your plans be dark and impenetrable as night, and when you move, fall like a thunderbolt."

~Sun Tzu, The Art of War

7

Conclusions and Future Perspectives

In this dissertation, eight major contributions in the form of four contributory chapters are presented. Extensive qualitative and quantitative results for each contribution are presented. The proposed models have achieved State-of-the-art (SOTA) results for 18 benchmark datasets. The summary of each contribution for four Vision task (VT)s (Semantic Segmentation (SS), Change Detection (CD), Image Translation (IT), Unsupervised Domain Adaptation (UDA)) is presented in this chapter.

7.1 Summary of Contributions

7.1.1 Semantic Segmentation

In the first contributory chapter, details of two proposed deep models named Aggregated Context Network (ACNet) and HybridNet for SS are presented. ACNet utilizes attention-based skip connection, dilated convolution extracted multi-context, and graph convolution-based far pixel relation mapping with hard negative mining for enhanced semantic segmentation. ACNet achieved $\sim$ 2-3 % improvement over the results produced by previous methods. HybridNet further improved the results for semantic segmentation with combined global pixel interaction, the dense dilated aggregated field of view, through voting and self-attention. HybridNet achieved state-of-the-art local and benchmark evaluation results with improvements of 0.7 % and 0.2 % in F1 Score(F_1) and Overall Accuracy (OA) values, respectively, for benchmark evaluation of the $Vaihingen$ dataset. For the $Potsdam$ dataset, it improved 0.04% and 0.03%, respectively, in F_1 and OA values from the previous results.

7.1.2 Change Detection

In the second chapter of this dissertation, two contributions named Difference image Reconstruction enhanced Multiresolution Network (DRMNet) and Triad Multitask Learning for change detection (TMLNet) are presented. DRMNet is proposed for efficient long-range dependency mapping and enhanced feature representation with multitask learning for effective change detection. DRMNet achived SOTA results for CDD, $SYSU$, and $LEVIRCD$ datasets. We also have set benchmark results for the $BCDD$ dataset for future comparison. To further improve the performance on change detection, TMLNet is proposed with triad reconstruction, enhanced backbone, multi-context local self-attention, and multitask learning. TMLNet achieved SOTA results on seven benchmark datasets with improvement on previous SOTA results set by DRMNet for four benchmark datasets.

7.1.3 Image Translation

In the third contributory chapter, details of proposed models (Encoder Decoder based Conditional GAN (EDCGAN), Multi Context Dense Network (MCDNet)) for IT. EDCGAN is an encoder-decoder-based model, for image translation from SAR to RGB with a multi-scale attentive discriminator for IT. EDCGAN attained SOTA results on benchmark dataset for three quantitative parameterise. MCDNet is proposed with the use of multiple spatial-space, self-attention, and dense residual attention for multi-frame super-resolution. SOTA results have been achieved by MCDNet on two benchmark datasets for two quantitative parameters.

7.1.4 Unsupervised Domain Adaptation

In the final contributory chapter, we presented two UDA techniques for SS and OT. Masked Domain Adversarial Adaptation Network (MDAANet) is proposed with masked domain dual adaptation, joint adversarial adaptation, domain reconstruction consistency, and feature dissimilarity alignment for domain adaptive semantic segmentation. MDAANet attained SOTA results on five benchmark domain adaptation tasks with performance improvement in the $\sim$ 1.5-10% range from previous methods. Reconstruction Assisted Domain Adaptation (RADA) is proposed with static style transfer and multi-level adaptation for domain adaptive object tracking. RADA achieved SOTA results on two benchmark nighttime adaptation datasets with improvements in the range of 3.7% - 11.4%.

7.2 Future Works

The following are some additional directions in which the current dissertation study can be expanded:

- In SS work, Multi modal (MM) and Self Supervised Learning (SSL) based model development will be explored. MM inputs (RGB with Synthetic Aperture Radar (SAR))

can helps in all weather semantic mapping. SSL can reduce the over-dependency on large annotated segmentation datasets with better feature representation.

- In CD work, we will extend our models for multi class Semantic Change Detection (SCD). SCD based model provides changes with semantic information for each pixel.
- In IT work, denoising diffusion-based model development will be explored because of its ability to produce high-quality synthesis, stable training, and avoid mode collapse.
- In UDA, current models will be enhanced to handle universal domain adaptation, which can further improve knowledge transfer between domains without requiring source domain data.

References

[1] Dstl satellite imagery feature detection. [Online]. Available: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/overview

[2] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sensing*, vol. 12, no. 14, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/14/2207

[3] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sensing*, vol. 11, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/11/1369

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[5] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on arXiv:1505.04597 [cs.CV]). [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a

[6] K. Chaurasia, R. Nandy, O. Pawar, R. R. Singh, and M. Ahire, "Semantic segmentation of high-resolution satellite images using deep learning," *Earth Science Informatics*, vol. 14, pp. 1–10, 12 2021.

[7] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Y. Sun, Y. Tian, and Y. Xu, "Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning," *Neurocomputing*, vol. 330, pp. 297–304, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218313821

[9] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[10] Y. Liu, D. Minh Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery," *Remote Sensing*, vol. 9, no. 6, p. 522, May 2017. [Online]. Available: http://dx.doi.org/10.3390/rs9060522

[11] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94 – 114, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271620300149

[12] D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," 2017. [Online]. Available: https://www.dropbox.com/s/8poqiexj75ewdxc/0190.pdf

[13] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[15] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019.

[16] N. Audebert, B. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," 03 2017, pp. 180–196.

[17] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[18] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Q. Liu, M. Kampffmeyer, R. Jenssen, and A. B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6309–6320, 2020.

[20] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," *NIPS*, 2018.

[21] X. Li, L. Zhang, A. You, M. Yang, K. Yang, and Y. Tong, "Global aggregation then local distribution in fully convolutional networks," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 244. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0432-paper.pdf

[22] S. Liu, K. Gao, J. Qin, H. Gong, H. Wang, L. Zhang, and D. Gong, "SE2Net: semantic segmentation of remote sensing images based on self-attention and edge

enhancement modules," *Journal of Applied Remote Sensing*, vol. 15, no. 2, pp. 1 – 16, 2021. [Online]. Available: https://doi.org/10.1117/1.JRS.15.026512

[23] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1633–1644, 2018.

[24] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Semantic segmentation of multisensor remote sensing imagery with deep ConvNets and higher-order conditional random fields," *Journal of Applied Remote Sensing*, vol. 13, no. 1, pp. 1 – 23, 2019. [Online]. Available: https://doi.org/10.1117/1.JRS.13.016501

[25] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 36–43.

[26] Y. Chen, D. Ming, and X. Lv, "Superpixel based land cover classification of vhr satellite image combining multi-scale cnn and scale parameter estimation," *Earth Science Informatics*, vol. 12, no. 3, pp. 341–363, 4 2019, communicated by: H. Babaie.

[27] N. Samet, S. Hicsonmez, and E. Akbas, "Houghnet: Integrating near and long-range evidence for bottom-up object detection," in *European Conference on Computer Vision (ECCV)*, 2020.

[28] F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, and N. Navab, "Hough-cnn: Deep learning for segmentation of deep brain regions in mri and ultrasound," *Computer Vision and Image Understanding*, vol. 164, pp. 92–102, 2017, deep Learning for Computer Vision. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314217300620

[29] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi, "Semi-convolutional operators for instance segmentation," in *European Conference on Computer Vision*, 2018.

[30] C. R. Qi, O. Litany, K. He, and L. Guibas, "Deep hough voting for 3d object detection in point clouds," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9276–9285.

[31] A. Sheshkus, A. Ingacheva, V. Arlazarov, and D. Nikolaev, "Houghnet: Neural network architecture for vanishing points detection," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 844–849.

[32] K. Ding, C. Huo, Y. Xu, Z. Zhong, and C. Pan, "Sparse hierarchical clustering for vhr image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 577–581, 2015.

[33] W. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," 1980.

[34] M. K. Ridd and J. Liu, "A comparison of four algorithms for change detection in an urban environment," *Remote Sensing of Environment*, vol. 63, no. 2, pp. 95–100, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425797001120

[35] S. Marchesi and L. Bruzzone, "Ica and kernel ica for change detection in multispectral remote sensing images," in *2009 IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, 2009, pp. II–980–II–983.

[36] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, 2007.

[37] J. Im, J. R. Jensen, and J. A. Tullis, "Object-based change detection using correlation image analysis and image segmentation," *International Journal*

of Remote Sensing, vol. 29, no. 2, pp. 399–423, 2008. [Online]. Available: https://doi.org/10.1080/01431160601075582

[38] A. Lefebvre, T. Corpetti, and L. Hubert-Moy, "Object-oriented approach and texture analysis for change detection in very high resolution images," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 4, 2008, pp. IV – 663–IV – 666.

[39] Q. Ke and P. Zhang, "Mccrnet: A multi-level change contextual refinement network for remote sensing image change detection," *ISPRS International Journal of Geo-Information*, vol. 10, no. 9, 2021. [Online]. Available: https://www.mdpi.com/2220-9964/10/9/591

[40] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "Ads-net:an attention-based deeply supervised network for remote sensing image change detection," *International Journal of Applied Earth Observation and Geoinformation*, vol. 101, p. 102348, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243421000556

[41] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2020.

[42] X. Zhang, Y. Yue, W. Gao, S. Yun, Q. Su, H. Yin, and Y. Zhang, "Difunet++: A satellite images change detection network based on unet++ and differential pyramid," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[43] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7232–7246, 2020.

[44] X. Jiang, S. Xiang, M. Wang, and P. Tang, "Dual-pathway change detection network based on the adaptive fusion module," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[45] X. Hou, Y. Bai, Y. Li, C. Shang, and Q. Shen, "High-resolution triplet network with dynamic multiscale feature for change detection on satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 103–115, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621001246

[46] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[47] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021.

[48] A. Chouhan, A. Sur, and D. Chutia, "Drmnet: Difference image reconstruction enhanced multiresolution network for optical change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4014–4026, 2022.

[49] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.

[50] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[51] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2021.

[52] W. Zhao, X. Chen, X. Ge, and J. Chen, "Using adversarial network for multiple change detection in bitemporal remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[53] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/10/1662

[54] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2021.

[55] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9976–9992, 2019.

[56] K. Song and J. Jiang, "Agcdetnet:an attention-guided network for building change detection in high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4816–4831, 2021.

[57] G. Cheng, G. Wang, and J. Han, "Isnet: Towards improving separability for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[58] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.

[59] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From w-net to cdgan: Bitemporal change detection via deep learning techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1790–1802, 2020.

[60] P. Jian, K. Chen, and W. Cheng, "Gan-based one-class classification for remote-sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[61] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[62] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," 2021.

[63] Y. Zhao, P. Chen, Z. Chen, Y. Bai, Z. Zhao, and X. Yang, "A triple-stream network with cross-stage feature fusion for high-resolution image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[64] J. Xu, C. Luo, X. Chen, S. Wei, and Y. Luo, "Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity," *Remote Sensing*, vol. 13, no. 15, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/15/3053

[65] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[66] H. Du, Y. Zhuang, S. Dong, C. Li, H. Chen, B. Zhao, and L. Chen, "Bilateral semantic fusion siamese network for change detection from multitemporal optical remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[67] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 3, pp. 545–559, 2018.

[68] E. Kalinicheva, J. Sublime, and M. Trocan, "Neural network autoencoder for change detection in satellite image time series," in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2018, pp. 641–642.

[69] L. Bergamasco, F. Bovolo, and L. Bruzzone, "Unsupervised change detection using hierarchical convolutional autoencoder," in *Image and Signal Processing for Remote Sensing XXVI*, L. Bruzzone, F. Bovolo, and E. Santi, Eds., vol. 11533, International Society for Optics and Photonics. SPIE, 2020, pp. 78 – 88. [Online]. Available: https://doi.org/10.1117/12.2571313

[70] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[71] H. Demirel and G. Anbarjafari, "Satellite image resolution enhancement using complex wavelet transform," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 123–126, 2010.

[72] ——, "Discrete wavelet transform-based satellite image resolution enhancement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 1997–2004, 2011.

[73] J. Anger, T. Ehret, C. de Franchis, and G. Facciolo, "Fast and accurate multi-frame super-resolution of satellite images," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-1-2020, pp. 57–64, 08 2020.

[74] T. Kato, H. Hino, and N. Murata, "Double sparsity for multi-frame super resolution," *Neurocomputing*, vol. 240, pp. 115–126, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523121730334X

[75] M. Kawulok, P. Benecki, D. Kostrzewa, and L. Skonieczny, "Evolving imaging model for super-resolution reconstruction," ser. GECCO '18. New York, NY,

USA: Association for Computing Machinery, 2018, p. 284–285. [Online]. Available: https://doi.org/10.1145/3205651.3205676

[76] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019.

[77] C. Shin, S. Kim, and Y. Kim, "Satellite image target super-resolution with adversarial shape discriminator," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[78] J. Ma, L. Zhang, and J. Zhang, "Sd-gan: Saliency-discriminated gan for remote sensing image superresolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1973–1977, 2020.

[79] M. Mostafa and S. Ahmed, "Satellite imagery super-resolution using squeeze-and-excitation-based gan," *International Journal of Aeronautical and Space Sciences*, 06 2021.

[80] P. Wang and E. Sertel, "Multi-frame super-resolution of remote sensing images using attention-based gan models," *Knowledge-Based Systems*, vol. 266, p. 110387, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123001375

[81] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[82] M. Rifat Arefin, V. Michalski, P.-L. St-Charles, A. Kalaitzis, S. Kim, S. E. Kahou, and Y. Bengio, "Multi-image super-resolution for remote sensing using deep recurrent networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 816–825.

[83] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Deep learning for super-resolution of unregistered multi-temporal satellite images," in *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2019, pp. 1–5.

[84] ——, "Deepsum++: Non-local deep neural network for super-resolution of unregistered multitemporal images," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 609–612.

[85] F. Dorr, "Satellite image multi-frame super resolution using 3d wide-activation neural networks," *Remote Sensing*, vol. 12, no. 22, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/22/3812

[86] M. R. Ibrahim, R. Benavente, F. Lumbreras, and D. Ponsa, "3drrdb: Super resolution of multiple remote sensing images using 3d residual in residual dense blocks," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 322–331.

[87] S. Mei, R. Jiang, X. Li, and Q. Du, "Spatial and spectral joint super-resolution using convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4590–4603, 2020.

[88] T. An, X. Zhang, C. Huo, B. Xue, L. Wang, and C. Pan, "Tr-misr: Multiimage super-resolution based on feature fusion with transformers," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1373–1388, 2022.

[89] Y. Xiao, Q. Yuan, J. He, Q. Zhang, J. Sun, X. Su, J. Wu, and L. Zhang, "Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, p. 102731, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243422000575

[90] D. Valsesia and E. Magli, "Permutation invariance and uncertainty in multitemporal image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[91] M. Märtens, D. Izzo, A. Kri, and D. Cox, "Super-resolution of proba-v images using convolutional neural networks," *Astrodynamics*, vol. 3, pp. 387–402, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195791394

[92] M. Kawulok, T. Tarasiewicz, J. Nalepa, D. Tyrna, and D. Kostrzewa, "Deep learning for multiple-image super-resolution of sentinel-2 data," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3885–3888.

[93] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo, "Self-supervised multi-image super-resolution for push-frame satellite images," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1121–1131.

[94] J. Wang, K. Gao, Z. Zhang, C. Ni, Z. Hu, D. Chen, and Q. Wu, "Multisensor Remote Sensing Imagery Super-Resolution with Conditional GAN," *Journal of Remote Sensing*, vol. 2021, p. 9829706, Sep. 2021.

[95] M. Razzak, G. Mateo-Garcia, L. Gómez-Chova, Y. Gal, and F. Kalaitzis, "Multi-spectral multi-image super-resolution of sentinel-2 with radiometric consistency losses and its effect on building delineation," 2021. [Online]. Available: https://arxiv.org/abs/2111.03231

[96] Q. Xu, X. Yuan, and C. Ouyang, "Class-aware domain adaptation for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[97] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation

of satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1067–1081, 2021.

[98] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 3816–3828, 2021.

[99] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2021.

[100] Z. Zhang, K. Doi, A. Iwasaki, and G. Xu, "Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 4, pp. 746–750, 2021.

[101] E. Arnaudo, A. Tavera, C. Masone, F. Dominici, and B. Caputo, "Hierarchical instance mixing across domains in aerial segmentation," *IEEE Access*, vol. 11, pp. 13 324–13 333, 2023.

[102] Q. Zhao, S. Lyu, B. Liu, L. Chen, and H. Zhao, "Self-training guided disentangled adaptation for cross-domain remote sensing image semantic segmentation," 2023.

[103] Y. Liu, X. Kang, Y. Huang, K. Wang, and G. Yang, "Unsupervised domain adaptation semantic segmentation for remote-sensing images via covariance attention," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[104] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[105] W. Liu, J. Liu, Z. Luo, H. Zhang, K. Gao, and J. Li, "Weakly supervised high spatial resolution land cover mapping based on self-training with weighted pseudo-labels,"

International Journal of Applied Earth Observation and Geoinformation, vol. 112, p. 102931, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843222001297

[106] L. Wang, P. Xiao, X. Zhang, and X. Chen, "A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4109–4121, 2023.

[107] H. Chen, L. Cheng, N. Li, Y. Yao, J. Cheng, and K. Zhang, "Sgda: An unsupervised domain adaptive semantic segmentation framework considering both imaging scene and geometric domain shifts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[108] A. Ma, C. Zheng, J. Wang, and Y. Zhong, "Domain adaptive land-cover classification via local consistency and global diversity," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[109] D. Zhao, B. Yuan, Y. Gao, X. Qi, and Z. Shi, "Ugcnet: An unsupervised semantic segmentation network embedded with geometry-consistency for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2021.

[110] Y. Cai, Y. Yang, Q. Zheng, Z. Shen, Y. Shang, J. Yin, and Z. Shi, "Bifdanet: Unsupervised bidirectional domain adaptation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 14, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/1/190

[111] L. Yan, B. Fan, S. Xiang, and C. Pan, "Cmt: Cross mean teacher unsupervised domain adaptation for vhr image semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[112] Y. Chen, X. Ouyang, K. Zhu, and G. Agam, *Semantic Segmentation in Aerial Images Using Class-Aware Unsupervised Domain Adaptation*. New York, NY,

USA: Association for Computing Machinery, 2021, p. 9–16. [Online]. Available: https://doi.org/10.1145/3486635.3491069

[113] M. Toldo, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings," in *Winter Conference on Applications of Computer Vision (WACV)*, January 2021.

[114] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7920–7930, 2020.

[115] M. Liu, P. Zhang, Q. Shi, and M. Liu, "An adversarial domain adaptation framework with kl-constraint for remote sensing land cover classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[116] X. Deng, H. L. Yang, N. Makkar, and D. Lunga, "Large scale unsupervised domain adaptation of segmentation networks with adversarial learning," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 4955–4958.

[117] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2022.

[118] J. Chen, G. Chen, B. Fang, J. Wang, and L. Wang, "Class-aware domain adaptation for coastal land cover mapping using optical remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 800–11 813, 2021.

[119] H. Ni, Q. Liu, H. Guan, H. Tang, and J. Chanussot, "Category-level assignment for cross-domain semantic segmentation in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[120] F. Zhang, Y. Shi, Z. Xiong, and X. X. Zhu, "Pseudo features guided self-training for domain adaptive semantic segmentation of aerial images," in *2023 Joint Urban Remote Sensing Event (JURSE)*, 2023, pp. 1–4.

[121] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[122] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, "Phase consistent ecological domain adaptation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[123] Z. Wu, X. Han, Y.-L. Lin, M. G. Uzunbas, T. Goldstein, S. N. Lim, and L. S. Davis, "Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 535–552.

[124] H. Gao, Y. Zhao, P. Guo, Z. Sun, X. Chen, and Y. Tang, "Cycle and self-supervised consistency training for adapting semantic segmentation of aerial images," *Remote Sensing*, vol. 14, no. 7, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/7/1527

[125] Y. Zhao, P. Guo, Z. Sun, X. Chen, and H. Gao, "Residualgan: Resize-residual dualgan for cross-domain remote sensing images semantic segmentation," 2022.

[126] M. Luo and S. Ji, "Cross-spatiotemporal land-cover classification from vhr remote sensing images with deep learning based domain adaptation," *ISPRS Journal*

of Photogrammetry and Remote Sensing, vol. 191, pp. 105–128, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271622001848

[127] W. Liu, F. Su, X. Jin, H. Li, and R. Qin, "Bispace domain adaptation network for remotely sensed semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[128] X. Deng, Y. Zhu, Y. Tian, and S. Newsam, "Scale aware adaptation for land-cover classification in remote sensing imagery," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 2160–2169.

[129] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level context memory," 2022. [Online]. Available: https://arxiv.org/abs/2208.07722

[130] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2021.

[131] A. Zheng, M. Wang, C. Li, J. Tang, and B. Luo, "Entropy guided adversarial domain adaptation for aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021.

[132] D. Wittich and F. Rottensteiner, "Appearance based deep domain adaptation for the classification of aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 82–102, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621002045

[133] X. Chen, S. Pan, and Y. Chong, "Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[134] L. Yan, B. Fan, S. Xiang, and C. Pan, "Adversarial domain adaptation with a domain similarity discriminator for semantic segmentation of urban areas," *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1583–1587, 2018.

[135] W. Liu and F. Su, "Unsupervised adversarial domain adaptation network for semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1978–1982, 2020.

[136] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 747–756. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w11/Tasar_StandardGAN_Multi-Source_Domain_Adaptation_for_Semantic_Segmentation_of_Very_High_CVPRW_2020_paper.html

[137] J. Guo, J. Yang, H. Yue, X. Liu, and K. Li, "Unsupervised domain-invariant feature learning for cloud detection of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2021.

[138] P. Dias, Y. Tian, S. Newsam, A. Tsaris, J. Hinkle, and D. Lunga, "Model assumptions and data characteristics: Impacts on domain adaptation in building segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[139] J. Lin, K. Yuan, R. Ward, and Z. J. Wang, "Xnet: Task-specific attentional domain adaptation for satellite-to-aerial scene," *Neurocomputing*, vol. 406, pp. 215–223, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220306020

[140] L. Yan, B. Fan, H. Liu, C. Huo, S. Xiang, and C. Pan, "Triplet adversarial domain adaptation for pixel-level classification of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3558–3573, 2020.

[141] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[142] S. Guo, T. Pridmore, Y. Kong, and X. Zhang, "An improved hough transform voting scheme utilizing surround suppression," *Pattern Recognition Letters*, vol. 30, no. 13, pp. 1241–1252, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865509001032

[143] T. Wollmann and K. Rohr, "Deep residual hough voting for mitotic cell detection in histopathology images," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 341–344.

[144] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogram. and Remote Sensing.*, vol. 145, pp. 78–95, 2018.

[145] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 96 – 107, 2018, deep Learning RS Data. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618300261

[146] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 156, pp. 1 – 13, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271619301741

[147] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158 – 172, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092427161630572X

[148] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017.

[149] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, 2017.

[150] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 06 2016.

[151] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[152] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham: Springer International Publishing, 2018, pp. 3–11.

[153] E. Kalinicheva, D. Ienco, J. Sublime, and M. Trocan, "Unsupervised change detection analysis in satellite image time series using deep learning combined with graph-based

approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1450–1466, 2020.

[154] Z. Wu, Z. Hu, and Q. Fan, "Superpixel-Based Unsupervised Change Detection Using Multi-Dimensional Change Vector Analysis and Svm-Based Classification," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I7, pp. 257–262, Jul. 2012.

[155] H. Zhang, M. Lin, G. Yang, and L. Zhang, "Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[156] Y. Jiao, Y. Niu, L. Liu, G. Zhao, G. Shi, and F. Li, "Dynamic range reduction of sar image via global optimum entropy maximization with reflectivity-distortion constraint," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2526–2538, 2018.

[157] b. Zhang, C. Wang, H. Zhang, and F. Wu, "An adaptive two-scale enhancement method to visualize man-made objects in very high resolution sar images," *Remote Sensing Letters*, vol. 6, pp. 725–734, 09 2015.

[158] X. Zhou and C. Zhang, "A perceptive uniform pseudo-color coding method of sar images," 11 2006, pp. 1 – 4.

[159] Z. Li, J. Liu, and J. Huang, "Dynamic range compression and pseudo-color presentation based on retinex for sar images," in *2008 International Conference on Computer Science and Software Engineering*, vol. 6, 2008, pp. 257–260.

[160] Q. Deng, Y. Chen, W. Zhang, and J. Yang, "Colorization for polarimetric sar image based on scattering mechanisms," in *2008 Congress on Image and Signal Processing*, vol. 4, 2008, pp. 697–701.

[161] W. Kang, Y. Xiang, F. Wang, L. Wan, and H. You, "Flood detection in gaofen-3 sar images via fully convolutional networks," *Sensors*, vol. 18, p. 2915, 09 2018.

[162] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 1, pp. 321–333, 2019.

[163] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, "Conditional image-to-image translation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[164] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[165] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arxiv*, 2016.

[166] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "Sar-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129 136–129 149, 2019.

[167] W. Zhang and M. Xu, "Translate sar data into optical image using ihs and wavelet transform integrated fusion," *Journal of the Indian Society of Remote Sensing*, vol. 47, 10 2018.

[168] X. Niu, D. Yang, K. Yang, H. Pan, Y. Dou, and F. Xia, "Image translation between high-resolution optical and synthetic aperture radar (sar) data," *International Journal of Remote Sensing*, vol. 42, no. 12, pp. 4758–4784, 2021. [Online]. Available: https://doi.org/10.1080/01431161.2020.1836426

[169] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220–1224, 2019.

[170] T. Zhang, F. Gao, J. Dong, and Q. Du, "Remote sensing image translation via style-based recalibration module and improved style discriminator," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[171] K. Enomoto, K. Sakurada, W. Wang, N. Kawaguchi, M. Matsuoka, and R. Nakamura, "Image translation between sar and optical imagery with generative adversarial nets," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 1752–1755.

[172] Q. Zhang, X. Liu, M. Liu, X. Zou, L. Zhu, and X. Ruan, "Comparative analysis of edge information and polarization on sar-to-optical translation based on conditional generative adversarial networks," *Remote Sensing*, vol. 13, no. 1, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/1/128

[173] J. Zhang, J. Zhou, and X. Lu, "Feature-guided sar-to-optical image translation," *IEEE Access*, vol. 8, pp. 70 925–70 937, 2020.

[174] J. Zhang, J. Zhou, M. Li, H. Zhou, and T. Yu, "Quality assessment of sar-to-optical image translation," *Remote Sensing*, vol. 12, no. 21, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/21/3472

[175] L. E. Christovam, M. H. Shimabukuro, M. L. B. T. Galo, and E. Honkavaara, "Evaluation of sar to optical image translation using conditional generative adversarial network for cloud removal in a crop dataset," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2021, pp. 823–828, 2021. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B3-2021/823/2021/

[176] T. Yu, J. Zhang, and J. Zhou, "Conditional gan with effective attention for sar-to-optical image translation," in *2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC)*. Los

Alamitos, CA, USA: IEEE Computer Society, apr 2021, pp. 7–11. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CTISC52352.2021.00009

[177] J. Guo, C. He, M. Zhang, Y. Li, X. Gao, and B. Song, "Edge-preserving convolutional generative adversarial networks for sar-to-optical image translation," *Remote Sensing*, vol. 13, no. 18, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/18/3575

[178] M. Zhang, C. He, J. Zhang, Y. Yang, X. Peng, and J. Guo, "Sar-to-optical image translation via neural partial differential equations," in *IJCAI*. ijcai.org, 2022, pp. 1644–1650.

[179] S. Fu, F. Xu, and Y. Jin, "Reciprocal translation between sar and optical remote sensing images with cascaded-residual adversarial networks," *Sci. China Inf. Sci.*, vol. 64, 2021.

[180] J. Zhao, W. Ni, Y. Zhou, Y. Chen, Z. Yang, and F. Bian, "Sar-to-optical image translation by a variational generative adversarial network," *Remote Sensing Letters*, vol. 13, no. 7, pp. 672–682, 2022. [Online]. Available: https://doi.org/10.1080/2150704X.2022.2068986

[181] X. Yang, J. Zhao, Z. Wei, N. Wang, and X. Gao, "Sar-to-optical image translation based on improved cgan," *Pattern Recognition*, vol. 121, p. 108208, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321003897

[182] K. Doi, K. Sakurada, M. Onishi, and A. Iwasaki, "Gan-based sar-to-optical image translation with region information," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2069–2072.

[183] G. Ji, Z. Wang, L. Zhou, Y. Xia, S. Zhong, and S. Gong, "Sar image colorization using multidomain cycle-consistency generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 2, pp. 296–300, 2021.

[184] P. Wang and V. M. Patel, "Generating high quality visible images from sar images using cnns," in *2018 IEEE Radar Conference (RadarConf18)*, 2018, pp. 0570–0575.

[185] M. Hakim, A. Ghazdali, and A. Laghrib, "A multi-frame super-resolution based on new variational data fidelity term," *Applied Mathematical Modelling*, vol. 87, pp. 446–467, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0307904X2030281X

[186] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "Stfnet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6552–6564, 2019.

[187] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep learning for multiple-image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1062–1066, 2020.

[188] N. Aburaed, A. Panthakkan, M. Al-Saad, M. C. E. Rai, S. A. Mansoori, H. Al-Ahmad, and S. Marshall, "Super-resolution of satellite imagery using a wavelet multiscale-based deep convolutional neural network model," in *Image and Signal Processing for Remote Sensing XXVI*, L. Bruzzone, F. Bovolo, and E. Santi, Eds., vol. 11533, International Society for Optics and Photonics. SPIE, 2020, p. 115331J. [Online]. Available: https://doi.org/10.1117/12.2573991

[189] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, "Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery," 2020. [Online]. Available: https://arxiv.org/abs/2002.06460

[190] L. Yan and K. Chang, "A new super resolution framework based on multi-task learning for remote sensing images," *Sensors*, vol. 21, no. 5, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1743

[191] D. Bull, N. Lim, and E. Frank, "Perceptual improvements for super-resolution of satellite imagery," in *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2021, pp. 1–6.

[192] M. Lee, I. Koo, K. Ko, and C. Kim, "Multi-image super-resolution via quality map associated temporal attention network," 02 2022.

[193] X. Liu, T. Feng, X. Shen, and R. Li, "Pmdrnet: A progressive multiscale deformable residual network for multi-image super-resolution of amsr2 arctic sea ice images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[194] T. Sun, Z. Xiong, Z. Wei, and Z. Wang, "Infrared image super-resolution method for edge computing based on adaptive nonlocal means," *The Journal of Supercomputing*, vol. 78, pp. 1–22, 04 2022.

[195] E. Mandanici, L. Tavasci, F. Corsini, and S. Gandolfi, "A multi-image super-resolution algorithm applied to thermal imagery," *Applied Geomatics*, vol. 11, 02 2019.

[196] H. Liu and Y. Gu, "Deep joint estimation network for satellite video super-resolution with multiple degradations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[197] M. Märtens, D. Izzo, A. Kržič, and D. Cox, "Super-resolution of proba-v images using convolutional neural networks," *Astrodynamics*, vol. 3, 08 2019.

[198] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7178–7193, 2020.

[199] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[200] J. Iqbal and M. Ali, "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 263–275, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301829

[201] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1144–1152.

[202] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 920–11 929.

[203] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6668–6677.

[204] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 549–12 556.

[205] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "Siamcar: Siamese fully convolutional classification and regression for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6269–6277.

[206] L. Lin*, H. Fan*, Z. Zhang, Y. Xu, and H. L. . equal contribution), "Swintrack: A simple and strong baseline for transformer tracking," in *Conf. on Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: https://arxiv.org/abs/2112.00995

[207] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *2022 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 13 598–13 608. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01324

[208] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13682. Springer, 2022, pp. 341–357. [Online]. Available: https://doi.org/10.1007/978-3-031-20047-2_20

[209] B. Li, C. Fu, F. Ding, J. Ye, and F. Lin, "ADTrack: Target-Aware Dual Filter Learning for Real-Time Anti-Dark UAV Tracking," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 1–8.

[210] Y. Sasagawa and H. Nagahara, "Yolo in the dark-domain adaptation method for merging multiple models," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 345–359.

[211] J. Ye, C. Fu, G. Zheng, Z. Cao, and B. Li, "Darklighter: Light up the darkness for uav tracking," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3079–3085.

[212] J. Ye, C. Fu, Z. Cao, S. An, G. Zheng, and B. Li, "Tracker meets night: A transformer enhancer for uav tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3866–3873, 2022.

[213] J. Zhu, H. Tang, Z.-Q. Cheng, J.-Y. He, B. Luo, S. Qiu, S. Li, and H. Lu, "Dcpt: Darkness clue-prompted tracking in nighttime uavs," *arXiv preprint arXiv:2309.10491*, 2023.

[214] L. Ma, D. Jin, N. An, J. Liu, X. Fan, Z. Luo, and R. Liu, "Bilevel fast scene adaptation for low-light image enhancement," *International Journal of Computer Vision*, pp. 1–19, 10 2023.

[215] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[216] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 769–15 778.

[217] J. Ye, C. Fu, G. Zheng, D. P. Paudel, and G. Chen, "Unsupervised Domain Adaptation for Nighttime Aerial Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1–10.

[218] J. Zhang, Z. Li, R. Wei, and Y. Wang, "Progressive domain-style translation for nighttime tracking," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 7324–7334. [Online]. Available: https://doi.org/10.1145/3581783.3612305

[219] L. Yao, H. Zuo, G. Zheng, C. Fu, and J. Pan, "Sam-da: Uav tracks anything at night with sam-powered domain adaptation," 2023.

[220] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2021, pp. 15 764–15 773. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01551

[221] K. Lu, C. Fu, Y. Wang, H. Zuo, G. Zheng, and J. Pan, "Cascaded denoising transformer for uav nighttime tracking," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3142–3149, 2023.

[222] C. Fu, T. Li, J. Ye, G. Zheng, S. Li, and P. Lu, "Scale-aware domain adaptation for robust uav tracking," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3764–3771, 2023.

[223] Y. Lv, W. Feng, S. Wang, G. Dauphin, Y. Zhang, and M. Xing, "Spectral-spatial feature enhancement algorithm for nighttime object detection and tracking," *Symmetry*, vol. 15, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2073-8994/15/2/546

[224] L. Sun, S. Kong, Z. Yang, D. Gao, and B. Fan, "Modified siamese network based on feature enhancement and dynamic template for low-light object tracking in uav videos," *Drones*, vol. 7, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2504-446X/7/7/483

[225] M. Kennerley, J. Wang, B. Veeravalli, and R. T. Tan, "2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2023, pp. 11 484–11 493. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01105

[226] J. Chen, Q. Sun, C. Zhao, W. Ren, and Y. Tang, "Rethinking unsupervised domain adaptation for nighttime tracking," in *Neural Information Processing*, B. Luo, L. Cheng, Z.-G. Wu, H. Li, and C. Li, Eds. Singapore: Springer Nature Singapore, 2024, pp. 391–404.

[227] X. Zheng, H. Cui, and X. Lu, "Multiple source domain adaptation for multiple object tracking in satellite video," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[228] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[229] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[230] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, Jul. 2017.

[231] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation." *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1802.html#abs-1802-02611

[232] M. A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4877–4885.

[233] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19.

[234] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. Torr, "Dual graph convolutional network for semantic segmentation," in *BMVC2019*.

[235] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[236] C. Zhang, J. Liu, F. Yu, S. Wan, Y. Han, J. Wang, and G. Wang, "Segmentation model based on convolutional neural networks for extracting vegetation from Gaofen-2 images," *Journal of Applied Remote Sensing*, vol. 12, no. 4, pp. 1 – 18, 2018. [Online]. Available: https://doi.org/10.1117/1.JRS.12.042804

[237] H. Bai, J. Cheng, X. Huang, S. Liu, and C. Deng, "Hcanet: A hierarchical context aggregation network for semantic segmentation of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[238] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, 2020.

[239] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[240] E. Del Barrio, J. A. Cuesta-Albertos, and C. Matrán, "An optimal transportation approach for assessing almost stochastic order," in *The Mathematics of the Uncertain*. Springer, 2018, pp. 33–44.

[241] D. Ulmer, C. Hardmeier, and J. Frellsen, "deep-significance-easy and meaningful statistical significance testing in the age of neural networks," *arXiv preprint arXiv:2204.06815*, 2022.

[242] R. Dror, S. Shlomov, and R. Reichart, "Deep dominance - how to properly compare deep neural models," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2773–2785. [Online]. Available: https://aclanthology.org/P19-1266

[243] N. Wambugu, Y. Chen, Z. Xiao, M. Wei, S. Aminu Bello, J. Marcato Junior, and J. Li, "A hybrid deep convolutional neural network for accurate land cover classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102515, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243421002221

[244] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[245] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-fpn for semantic segmentation of fine-resolution remotely sensed images," *International Journal of Remote Sensing*, vol. 43, no. 3, pp. 1131–1155, 2022. [Online]. Available: https://doi.org/10.1080/01431161.2022.2030071

[246] C. Hazırbaş, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," 11 2016.

[247] C. Zhang, W. Jiang, and Q. Zhao, "Semantic segmentation of aerial imagery via split-attention networks with disentangled nonlocal and edge supervision," *Remote Sensing*, vol. 13, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/6/1176

[248] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *Advances in Neural Networks – ISNN 2019*, H. Lu, H. Tang, and Z. Wang, Eds. Cham: Springer International Publishing, 2019, pp. 388–401.

[249] V. Jampani, D. Sun, M.-Y. Liu, M.-H. Yang, and J. Kautz, "Superpixel sampling networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[250] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region mutual information loss for semantic segmentation," in *NeurIPS*, 2019.

[251] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 2115–2118.

[252] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved unet++," *Remote Sensing*, vol. 11, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/11/1382

[253] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 214–217.

[254] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 266–270, 2019.

[255] Y. Zhang, S. Zhang, Y. Li, and Y. Zhang, "Coarse-to-fine satellite images change detection framework via boundary-aware attentive network," *Sensors*, vol. 20, no. 23, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/6735

[256] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301532

[257] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? roll the dice and demand attention," 2020.

[258] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural*

Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/3416a75f4cea9109507cacd8e2f2aefc-Paper.pdf

[259] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883.

[260] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change Detection in Remote Sensing Images Using Conditional Adversarial Networks," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 422, pp. 565–571, May 2018.

[261] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[262] H. Lee, K. Lee, J. H. Kim, Y. Na, J. Park, J. P. Choi, and J. Y. Hwang, "Local similarity siamese network for urban land change detection on remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4139–4149, 2021.

[263] P. Chen, D. Hong, Z. Chen, X. Yang, B. Li, and B. Zhang, "Fccdn: Feature constraint network for vhr image change detection," 2021.

[264] J. Baxter, "A bayesian/information theoretic model of learning to learn viamultiple task sampling," *Mach. Learn.*, vol. 28, no. 1, p. 7–39, jul 1997. [Online]. Available: https://doi.org/10.1023/A:1007327622663

[265] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[266] S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1716–1724. [Online]. Available: https://doi.org/10.1145/3123266.3123425

[267] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, Y. Zhang, A. Zhang, S. Lv, and B. Jiang, "S2looking: A satellite side-looking dataset for building change detection," 2021.

[268] X. Wu, L. Yang, Y. Ma, C. Wu, C. Guo, H. Yan, Z. Qiao, S. Yao, and Y. Fan, "An end-to-end multiple side-outputs fusion deep supervision network based remote sensing image change detection algorithm," *Signal Processing*, vol. 213, p. 109203, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168423002773

[269] Z. Wang, C. Peng, Y. Zhang, N. Wang, and L. Luo, "Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss," *Neurocomputing*, vol. 457, pp. 155–167, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221009838

[270] Y. Zhou, Y. Feng, S. Huo, and X. Li, "Joint frequency-spatial domain network for remote sensing optical image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[271] X. Zhang, M. Tian, Y. Xing, Y. Yue, Y. Li, H. Yin, R. Xia, J. Jin, and Y. Zhang, "Adhr-cdnet: Attentive differential high-resolution change detection network for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[272] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multiscale siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[273] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 15 173–15 182. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01491

[274] Z. Zheng, S. Tian, A. Ma, L. Zhang, and Y. Zhong, "Scalable multi-temporal remote sensing change data generation via simulating stochastic change process," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 21 818–21 827.

[275] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[276] J. Pan, W. Cui, X. An, X. Huang, H. Zhang, S. Zhang, R. Zhang, X. Li, W. Cheng, and Y. Hu, "Mapsnet: Multi-level feature constraint and fusion network for change detection," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, p. 102676, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243422000022

[277] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 702–716.

[278] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[279] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[280] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2868–2876.

[281] Z. Zuo and Y. Li, "A sar-to-optical image translation method based on pix2pix," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3026–3029.

[282] X. Yang, Z. Wang, J. Zhao, and D. Yang, "Fg-gan: A fine-grained generative adversarial network for unsupervised sar-to-optical image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[283] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Deepsum: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2019.

[284] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," 01 2021, pp. 1368–1377.

[285] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[286] L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, and S. Ouyang, "Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive

learning and adversarial learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[287] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 20–33, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621000423

[288] H. Huang, B. Li, Y. Zhang, T. Chen, and B. Wang, "Joint distribution adaptive-alignment for cross-domain segmentation of high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[289] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 7472–7481. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00780

[290] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," *arXiv preprint arXiv:2101.10979*, 2021.

[291] Y. Zhao, P. Guo, H. Gao, and X. Chen, "Depth-assisted residualgan for cross-domain aerial images semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[292] Y. Zheng, L. He, X. Wu, and C. Pan, "Self-training and multi-level adversarial network for domain adaptive remote sensing image segmentation," *Neural Process. Lett.*, vol. 55, no. 8, p. 10613–10638, aug 2023. [Online]. Available: https://doi.org/10.1007/s11063-023-11341-x

[293] X. Li, M. Luo, S. Ji, L. Zhang, and M. Lu, "Evaluating generative adversarial networks based image-level domain transfer for multi-source remote sensing image segmentation and object detection," *International Journal of Remote Sensing*, vol. 41, no. 19, pp. 7343–7367, 2020.

[294] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[295] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5369–5378, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52350875

[296] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, p. 211–252, dec 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y

Publications Related to Thesis

Journals

1. **A. Chouhan**, A. Sur, D. Chutia and Shiv Prasad Aggarwal, "TMLNet: Triad Multitask Learning for change detection", in Neurocomputing, Volume 599, 2024, 128080, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2024.128080. [Chapter 4]

2. **A. Chouhan**, A. Sur, D. Chutia and Shiv Prasad Aggarwal , "HybridNet: Integrating multiple approaches for aerial semantic segmentation" in SN Computer Science, 2023 [Chapter 3]

3. **A. Chouhan**, A. Sur and D. Chutia, "DRMNet: Difference Image Reconstruction Enhanced Multiresolution Network for Optical Change Detection," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 4014-4026, 2022, doi: 10.1109/JSTARS.2022.3174780. [Chapter 4]

Conferences

1. **A. Chouhan**, M. Chandak, A. Sur and D. Chutia, S.P. Aggarwal, 2025, RADA: Reconstruction Assisted Domain Adaptation for night time object tracking. Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science, vol 15310. Springer, Cham. https://doi.org/10.1007/978-3-031-78192-6_21 [Chapter 6]

2. **A. Chouhan**, H. Motwani, A. Sur and D. Chutia, S.P. Aggarwal, 2024, "MCDNet: Multi Context Dense Network for multi-frame super resolution

of satellite images", In Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '23). Association for Computing Machinery, New York, NY, USA, Article 33, 1–9. https://doi.org/10.1145/3627631.3627664 [Chapter 5]

3. **A. Chouhan**, A. Sur and D. Chutia, "Aggregated Context Network For Semantic Segmentation Of Aerial Images," 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 1526-1530, doi: 10.1109/ICIP46576.2022.9898016. [Chapter 3]

4. **A. Chouhan**, N. Jindal, A. Sur, D. Chutia, and S. P. Aggarwal. 2023. EDCGAN: Encoder Decoder based Conditional GAN for SAR to Optical image translation. In Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '22). Association for Computing Machinery, New York, NY, USA, Article 38, 1–8. https://doi.org/10.1145/3571600.3571639 [Chapter 5]

5. **A. Chouhan**, A. Agrawal and A. Sur, "Unsupervised Change Detection in Very High Resolution Multi-Spectral Images," 2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS), Ahmedabad, India, 2021, pp. 293-296, doi: 10.1109/InGARSS51564.2021.9791474. [Chapter 4]

Under Review

1. **A. Chouhan**, A. Sur, D. Chutia and S. P. Aggarwal, MDAAF: Masked Domain Adversarial Adaptation Framework [Chapter 6]

Publications Outside Thesis

1. Source Free Universal Domain Adaptation for Remote Sensing Image Classification [Manuscript prepared]

Department of Computer Science and Engineering

Indian Institute of Technology Guwahati

Guwahati 781039, India