भारतीय प्रौद्योगिकी संस्थान दिल्ली
हौज खास, नई दिल्ली-110 016, भारत
Indian Institute of Technology Delhi
Hauz Khas, New Delhi-110 016, India

Phone(O) +91 - 11 - 2659 1084
Facsimilie +91 - 11 - 2658 1606
E-Mail ID sumantra@ee.iitd.ac.in
sumantra.dutta.roy@gmail.com
http://www.cse.iitd.ac.in/∼sumantra

सुमन्त्र दत्त राय
प्राध्यपक, विद्युत प्रौद्योगिकी विभाग
Sumantra Dutta Roy
Professor, Department of Electrical Engineering

Date: 13th May, 2025

To: Prof. Bulu Pradhan,
Associate Dean of Academic Affairs (PG)
Academic Affairs Section, IIT Guwahati, Guwahati - 781039, Assam.
`oadoaapg@iitg.ac.in, adoaapg@iitg.ac.in, bulu@iitg.ac.in`

SUBJECT: Evaluation Report of the Ph. D. Thesis of Mr. Sandipan Sarma (Roll No. 196101005), email dated 04 March, 2025

Dear Professor Bulu Pradhan,

I enclose the evaluation report for the thesis of Mr. Sandipan Sarma (Roll No. 196101005), "Zero-Shot Learning for Visual Recognition and Detection Tasks," with this letter.

With warm regards,

Sumantra Dutta Roy

(Sumantra Dutta Roy)

encl: Evaluation report

Detailed report on the Ph.D. Thesis of Mr. Sandipan Sarma (Roll No. 196101005), "Zero-Shot Learning for Visual Recognition and Detection Tasks"

The volume of work done by Mr. Sandipan Sarma (Roll No. 196101005) is significant, and the novel ideas presented in the work make it a worthy candidate for the award of the degree of Doctor of Philosophy.

The thesis makes significant contributions to the knowledge in the area, all tied up with the common theme of zero-shot learning. The thesis has resulted in *two top journal papers*, with *one more under review in another high impact factor journal*. The candidate has papers in *top conferences: ICPR, ICASSP, ACM MM, BMVC* and has *one more paper in review in a top conference (ICIP)*. The candidate also has one publication outside the thesis: *an ICASSP paper*. All this is at the time of submission of the thesis! This is representative of the excellent quality of work in the thesis. *I strongly recommend the thesis for the award of the Ph. D. Degree, in its present form.*

Here are some specific points with regard to the thesis.

1. I find the thesis extremely well-motivated and well laid-out, which made reading the thesis an extremely enjoyable experience.
2. Chapter 1 introduces the area well, motivates the work, and summarises the contributions of the thesis, and gives its organisation.
3. Chapter 2 gives a concise and yet, pointed report on basic background concepts and related work in the area. The chapter first presents a summary of some concepts used across the thesis: Region-CNN, Contrastive Language-Image Pre-training (CLIP), and Detection Transformer (DETR). The chapter presents the literature survey in related areas: zero-shot learning, zero-shot object detection, zero-shot action recognition, zero-shot human-object interaction detection, and zero-shot underwater gesture recognition, which relate to the 5 contributory chapters of the thesis.
4. Chapter 3 considers the identification of diverse and rare training classes for zero-shot learning. The motivation for the work in this chapter is quite clear: the earlier breakthroughs in zero-shot learning with generative models is governed by the frequency of the seen class samples, and the bias thereof.
   Given an attribute-based dataset, the proposed method can intelligently yield the suitable 'seen classes' for training zero-shot models. The work in the chapter also seeds a visual semantic mining algorithm for class acquisition capturing both the diversity as well as rarity in the object domain. Fig.3.3 gives a beautiful workflow diagram of the work in this chapter, which gives a good visual representation of the two main contributions of this chapter. The chapter shows impressive results on large *de facto* standard datasets.
5. Chapter 4 works on resolving semantic confusions for improved zero-shot detection. The work in this chapter is also well-motivated. Generative techniques such as GANs can generate unseen-class samples based on the semantics of the GAN trained on seen-class data. However, models are often unable to distinguish between semantically similar classes.

The chapter proposes a generative model with a triplet loss and a cycle consistency loss. The triplet loss incorporates the degree of dissimilarity between classes, and incorporates them in generated samples. The cycle consistency loss ensures fidelity of the generated samples with their own semantics. The architecture incorporates features from a Faster RCNN and a pre-trained ResNet-101 for feature extraction. The feature synthesis is via a conditional Wasserstein GAN. Fig.4.1 gives a nice overview of the proposed model.

6. Chapter 5 models multi-scale local context and action relationships for zero-shot action recognition. The chapter starts with an interesting observation for a topical zone of research namely, action recognition. The chapter identifies that existing research does not model the local spatial context of objects and action environments.
   The chapter proposes a novel two-stage framework (Fig.5.3) involving a local context aggregating temporal transformer (this takes image and text encodings from a pre-trained Image-based Vision-Language (I-VL) model), and a graph attention network. The local context aggregating temporal transformer captures multi-scale local context with dilated convolutional layers, and produces suitable video embeddings. The graph attention network models semantic relationships between classes, in tune with the video embeddings produced by the transformer part.
   The training phase involves the image encoder of a pre-trained Contrastive Image-Language Pre-training (CLIP) system. This produces frame-wise video encodings. The proposed method feeds a suitable prompt to the text encoder for a text encoding (the semantics). The visual branch with the temporal transformer captures multi-scale local context and aggregates them from the temporal dimension. The semantic branch models a node's relationships with its neighbours via a multi-head attention mechanism, and learns seen-unseen associations.
   The two-stage training trains the temporal transformer part with seen-class data alone, and is then used as a classifier. These outputs are aligned with those from a trained graph attention module.

7. Chapter 6 explores top-down perceptions for zero-shot human-object interaction (HOI) detection. The chapter starts from the premise that the combinatorial and hence, exponential number of object-action combinations lead to a long-tailed distribution problem. The chapter recognises that for better scene interpretation, human-object interaction cues must be incorporated in the encoder stage itself.
   The chapter proposes a human psychology-inspired framework (Fig.6.2): we humans tend to grasp well-defined concepts first, and then associate them with abstract concepts during scene understanding. The well-defined concepts involve object detection. The actions are the abstract concepts associated with them. The architecture is based on a detection transformer (DETR), and CLIP.
   The chapter proposes an asymmetric co-attention mechanism utilising multimodal information with zero-shot capabilities. This jointly learns how visual understanding is influenced by textual knowledge about the few nominated objects, and vice versa. The object-probed feature map gives conditional information to associate likely actions. The chapter re-designs the basic focal loss to focus on 'hard examples', a related-ness factor, and an inter-class dependency factor.

8. Chapter 7 deals with a novel area of research, zero-shot underwater gesture recognition. The work is again topical, since an underwater gesture system has been devised for divers. The domain difficulties result in unseen gestures causing the system to fail.

The chapter proposes new seen-unseen splits for gesture classes in a *de facto* standard dataset. The chapter proposes a novel two-stage framework (gated cross-attention transformer, and a conditional Wasserstein GAN, Fig.7.3). The former learns visual gestures cues and feeds them to the GAN which learns to mimic the feature distribution. The GAN is the synthesiser for unseen classes (zero-shot learning). The classifier is trained with seen class data, and synthesised unseen class data, in the zero-shot learning.

9. Chapter 8 concludes the thesis, and gives some pointers for future work in the area.

10. The thesis has some extremely minor typographical errors which can be corrected very easily by the candidate, on a close and careful observation of the typed text.
    This can be done when the final thesis is submitted to the Institute. This is not hair-splitting: I would like to see this large volume of excellent research be as perfect as possible. Since these are absolutely minor, and in no way take the sheen off the academic contributions of the thesis, I have recommended the first option, *The thesis is commended in its present form, for award of the Ph.D. degree.*

    (a) Any one format can be followed consistently for the references, such as the IEEE format, an Elsevier format, and so on.

    (b) There are some issues with suitable capitalisation. There is some inconsistent capitalisation in the name of the journal or conference, and the title of the paper. In the title, all important words should be capitalised in its first character, or none at all, except the first: this should be consistent across all references.

    (c) Some *de facto* standard abbreviations and words should be suitably capitalised. Examples include 'Covid-19' in [1], [169], 'CNN' in [2], [47], 'CBAM' in [7], 'YOLO' in [19], 'ImageNet' in [22], [50], 'GTNet' in [29], 'Pascal' and 'VOC' in [32], 'UCF' in [33], 'X-CAUNet' in [36] (the candidate's own paper!), 'ALIP' in [39], 'ActivityNet' in [40], 'Kinetics' in [41] and [42], 'HRI' in [45], 'DeViSE' in [59], 'HSVA' in [72], 'MSDN' in [77], 'GCN' in [81], 'NUS' in [89], 'YOLO' in [96], 'SIR' in [101], 'VDARN' in [106], 'GloVe' in [113], 'BERT' in [121], 'GAN' in [122], [175], [177], [178], 'ViViT' in [134], 'ConsNet' in [140], 'ZSGL' in [180], 'MTS/IEEE' in [155], 'DARE' in [160], 'MobileNets' in [163], 'WordNet' in [165], 'AUVs' in [170], 'CI-GNN' in [212], 'fMRI' in [218], 'QPIC' in [224], 'ERNet' in [229], 'FGAHOI' in [230] ('HOI' should be written as such, unlike what has been mentioned across a very large number of references), 'ICAN' in [232], 'IPGN' in [237], 'ACP++' in [238], 'DRG' in [239], 'PPDM++' in [240], 'DSSF' in [241] and 'GELU' in [251].

    (d) Words based on the names of countries and places should have their first character capitalised. An example of the same is 'Hungarian' in [54], 'Singapore' in [89], 'English' in [165] and 'Siamese' in [222].

    (e) It is a convention for non-English words to be printed in an italicised manner. Examples of such errors include *quo vadis* in [99].

    (f) A few references are incomplete, such as [30], [78], [183], [204] and [212].

    (g) The same journal or conference name appears differently at different places. An example of the same is in [4] and [7], and [14] and [18]. The use of BibTeX macros can remove these inconsistencies, if one uses LaTeX as the typesetting tool.

    (h) There should be uniformity in conference publications, with the same number of fields, as far as possible. For instance, the following format is widely accepted:

30, 78, 183 are book chapters, etc. - so no more info about them present. 204 looks fine, rectified 212

too troublesome, can't change them all, even with macros

- M. Izzetoglu, B. Onaral, N. Bilgutay. "Wavelet domain least squares deconvolution for ultrasonic backscattered signals." In *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 1, pp. 321 – 324, 2000.

problematic with bib style - couldn't do

One understands that it is quite difficult to have a consistent referencing for conferences, it is nice to try to have the same minimal set of parameters for all conferences, as far as is practically possible. The same goes for journals, books and book chapters as well: have a minimal set of parameters for each type.

(i) The general convention of using *et al.* is to have the first author's name followed by *et al.* for a publication with three or more authors. References in error include [3], [13], [19], [41], [50], [51], [53] and [170].

(j) The set of publications of the candidate (p.202) should be put in the same format as the other references.

I emphasise that this is not hair-splitting: I would like to see this large volume of excellent research be as perfect as possible.

Sumantra Dutta Roy

13th May, 2025 (Sumantra Dutta Roy)