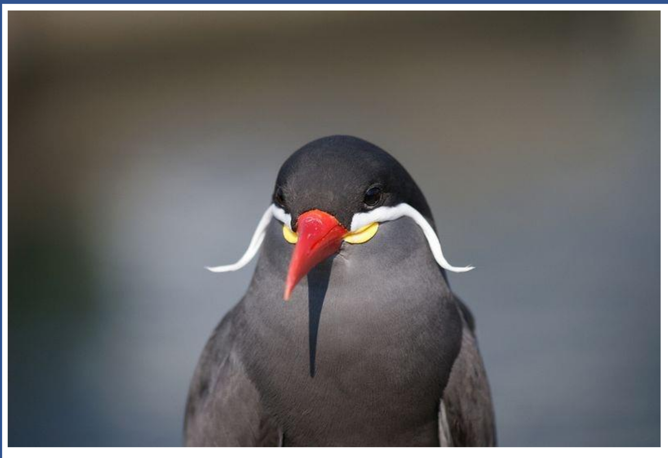
Resolving Semantic Confusions for Improved Zero-Shot Detection

BMVC 2022

Sandipan Sarma, Sushil Kumar, Arijit Sur
Department of CSE, Indian Institute of Technology Guwahati, India

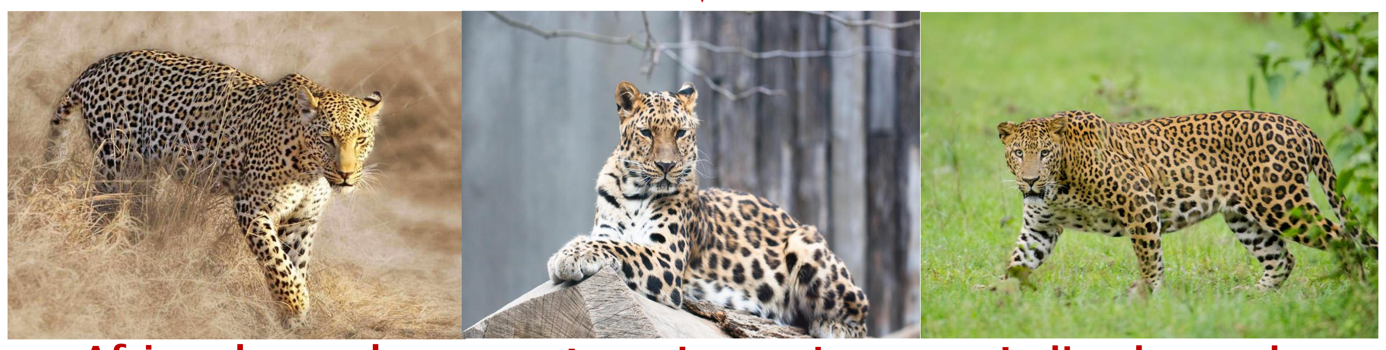
Motivation



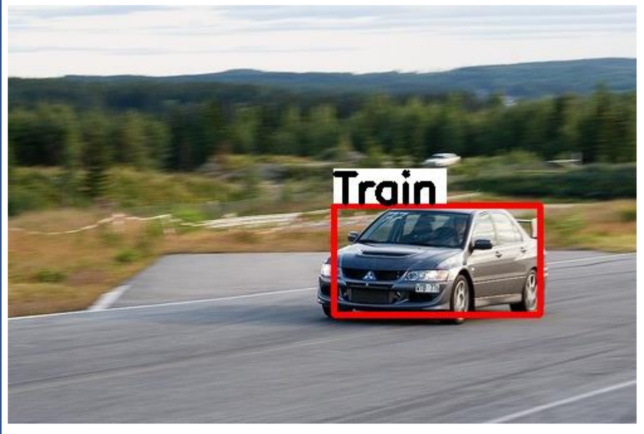
Zero-shot Detection (ZSD) models can alleviate all these issues

Impossible to collect images of all possible objects, especially rare, for training supervised object detection models

Supervised detection models cannot adapt to evolving object appearances, like futuristic cars

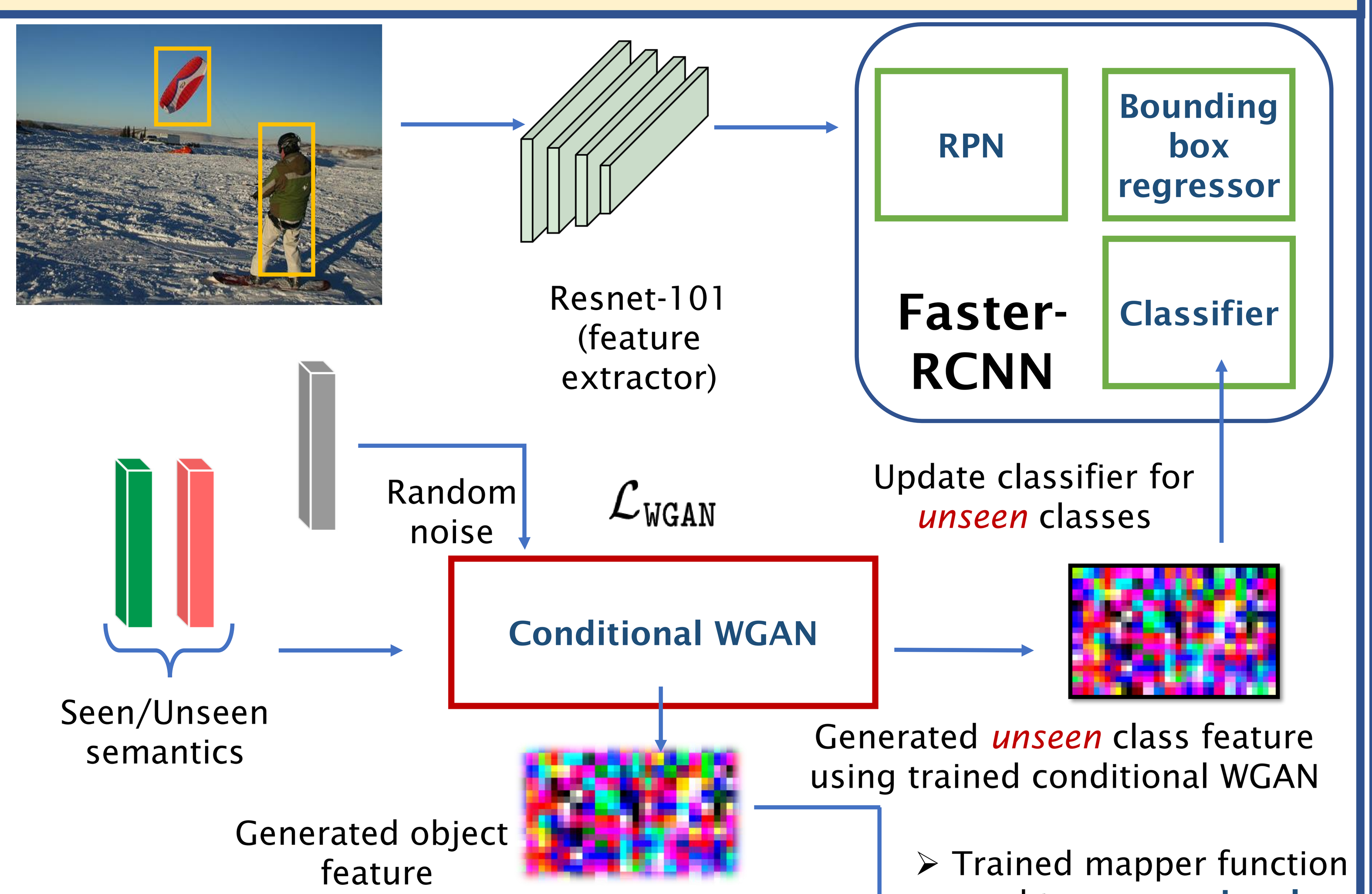


Fine-grained annotations time-consuming and require expert analysis

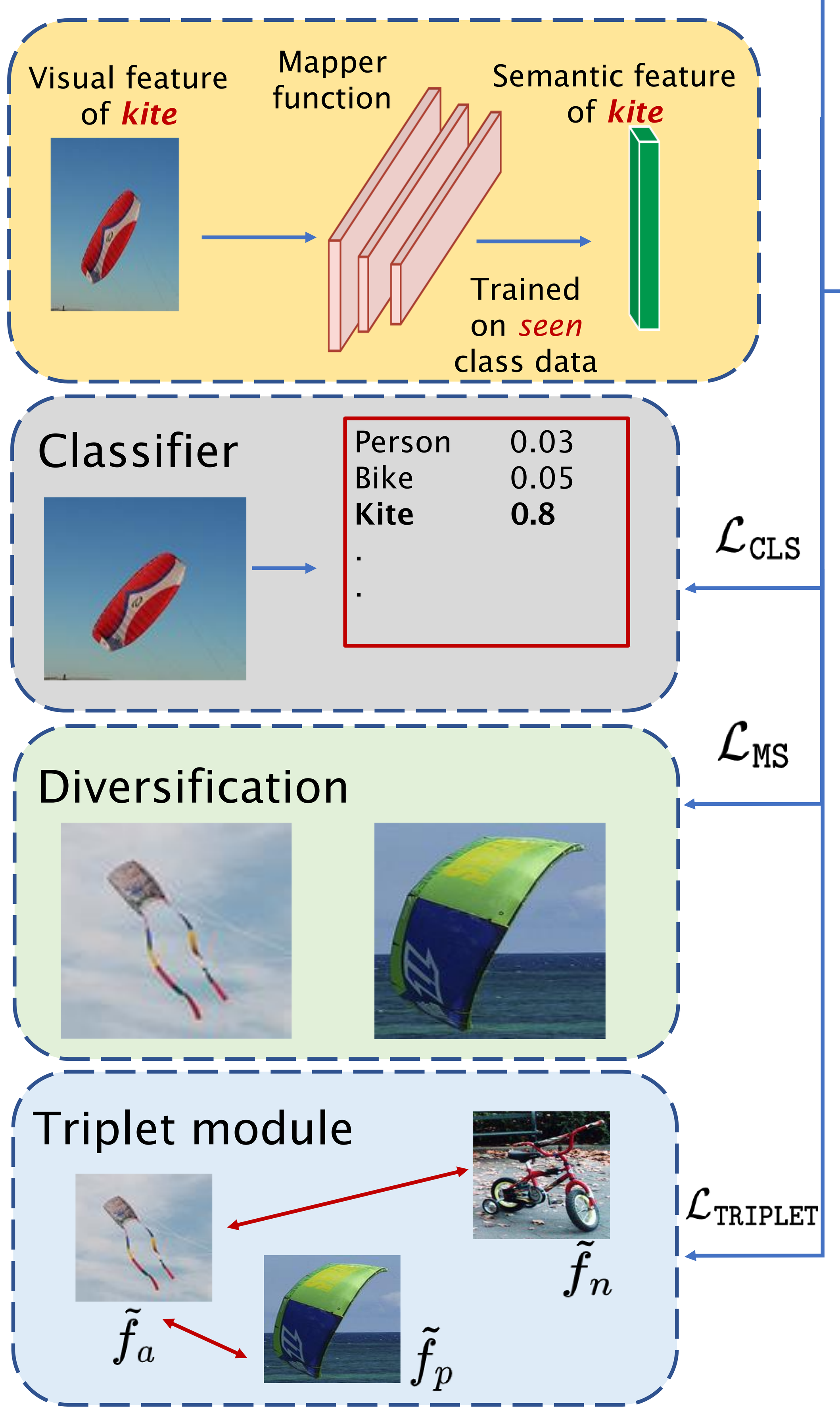


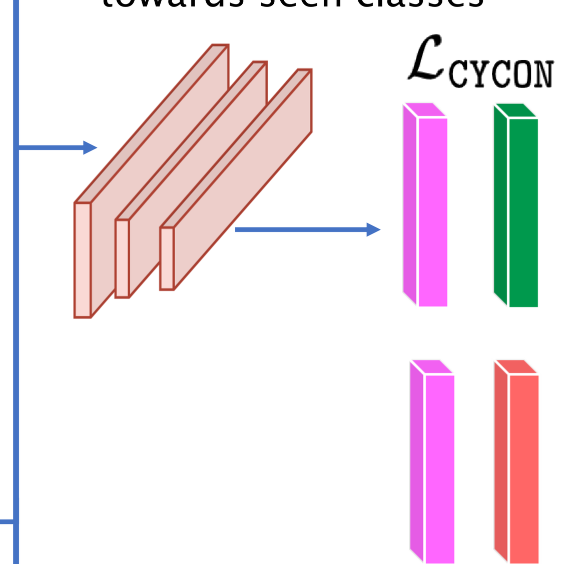
- **Semantic Confusion**: Knowledge transfer in existing ZSD models not discriminative enough to differentiate between objects with similar attributes (*semantics*) like *car* and *train*
- Low average precisions can misguide applications like marine debris detection

Proposed Framework



- Trained mapper function used to ensure **visual-semantic consistency**
- Cyclic-consistency loss reduces bias of generated features towards seen classes





- Classification loss encouraging **discriminative feature generation**
- Mode-seeking loss **prevents** GAN from falling into a **mode collapse**
- Incorporating **inter-class dissimilarity** within the generated object features
- Margin of dissimilarity **(Δ) computed** on the basis of mean and variance of class semantics

Experiments

Datasets

COCO – Common Objects in Context

- **80 object categories** with bounding-box annotations
- **Seen/Unseen classes** = 65/15
- **Train/test images** = 62,300/10,098

PASCAL2 – Pattern Analysis, Statistical Modelling and Computational Learning

- **20 object categories** with bounding-box annotations
- **Seen/Unseen classes** = 16/4
- **Train/test images** = 5.981/1,402 + 4,836

Source of inputs

fastText – Library for efficient text classification and representation learning

- **Attribute (Semantic) source**
- **300-dimensional** semantic vector per object class

IMAGENET

- **Visual feature source**
- Pre-trained ResNet-101 trained on ImageNet, excluding classes common with *unseen* objects of MS-COCO and PASCAL-VOC, **as per zero-shot criterion**

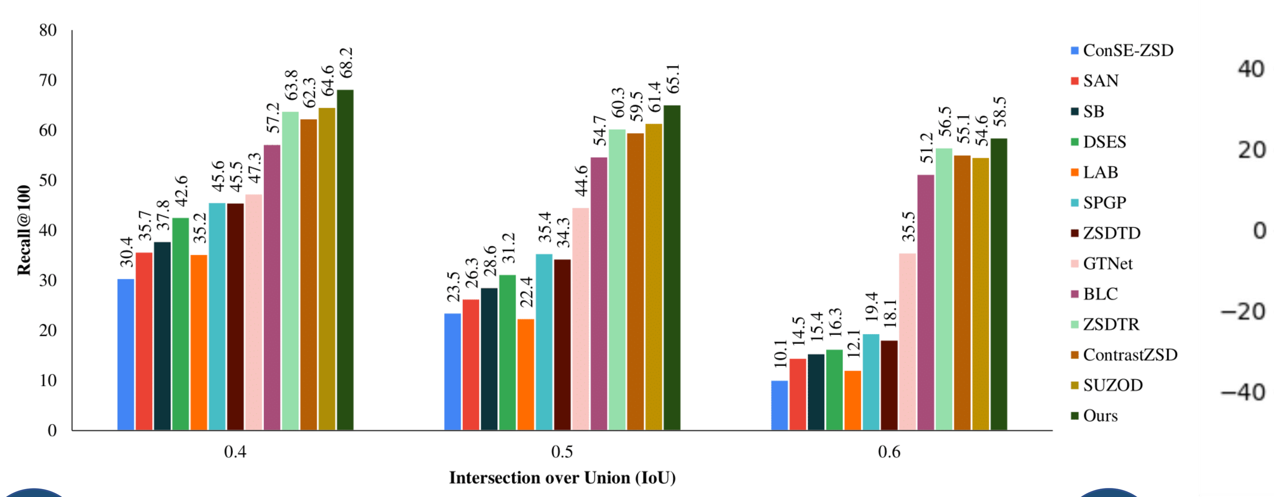
Results

1 Mean Average Precision (in %) with MS-COCO

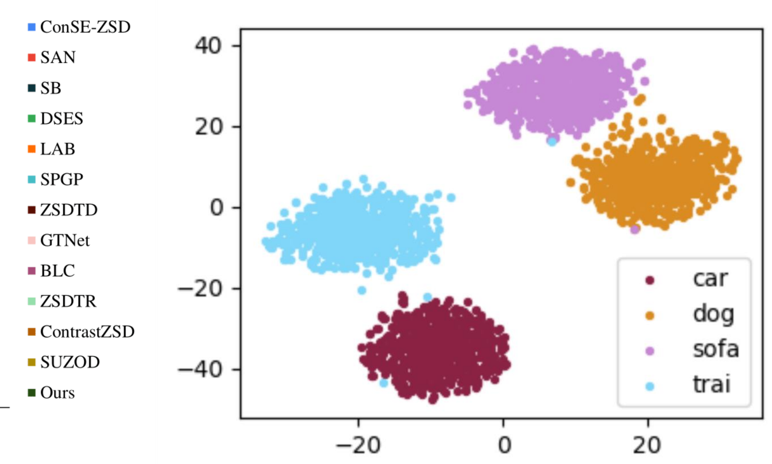
Best results (red) / Second-best results (blue)

Method	Year	ZSD	GZSD Seen	GZSD Unseen	GZSD HM
PL	2018	12.40	34.07	12.40	18.18
BLC	2020	14.70	36.00	13.10	19.20
ACS-ZSD	2020	15.34	-	-	-
SUZOD	2020	17.30	37.40	**17.30**	**23.65**
ZSDTR	2021	13.20	40.55	13.22	20.16
ContrastZSD	2022	**18.60**	**40.20**	16.50	23.40
Ours	2022	**20.10**	37.40	**20.10**	**26.15**

2 Recall@100 with MS-COCO



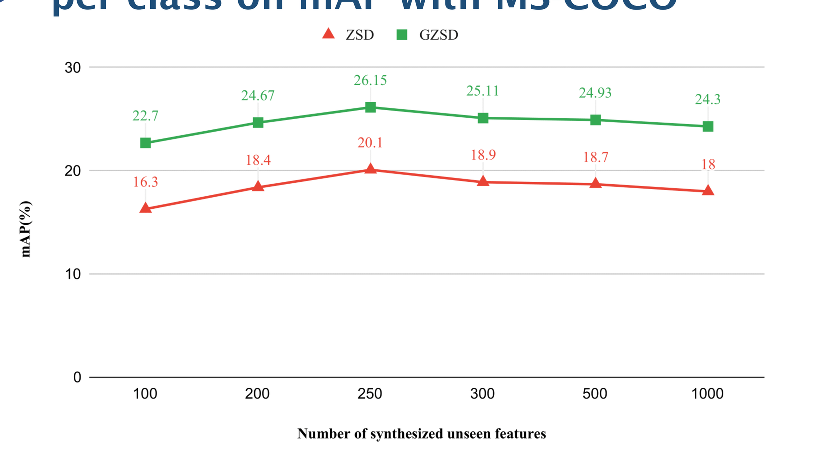
3 t-SNE plot of generated features with PASCAL-VOC



4 Effect of loss components on mAP (in %) with MS-COCO

$\mathcal{L}_{WGAN}$	$\mathcal{L}_{MS}$	$\mathcal{L}_{CLS}$	$\mathcal{L}_{CYCON}$	$\mathcal{L}_{TRIPLET}$	mAP
✓	✓	✓	✓		18.4
✓		✓	✓	✓	18.5
✓	✓		✓	✓	19.3
✓	✓	✓		✓	19.4
✓	✓	✓	✓	✓	**20.1**

5 Effect of number of generated features per class on mAP with MS-COCO



Takeaways

- Visual-feature based **triplet strategy** for addressing **inter-class dissimilarities**, along with **cyclic-consistency** between generated visual features of an object and its semantic vector, encourages a generative network to synthesize discriminative features, **reducing semantic confusion**
- **False-positive rate** and **misclassification of localized objects** (especially *unseen* objects) can be addressed as a by-product of resolving semantic confusion
- With conditional feature generation, performance usually **stays low** for *unseen* objects that are **semantically very different** from every *seen* object class

Acknowledgements

This work is supported by IITG Technology Innovation and Development Foundation (IITGTI&DF), which has been set up at IIT Guwahati as a part of the National Mission on Interdisciplinary Cyber Physical Systems (NMICPS). IITGTI&DF is undertaking research, development and training activities on Technologies for Under Water Exploration with the financial assistance from Department of Science and Technology, India through grant number DST/NMICPS/TIH12/IITG/2020. Authors gratefully acknowledge the support provided for the present work. We also acknowledge the Department of Biotechnology, Govt. of India for the financial support for the project BT/COE/34/SP28408/2018 (for computing resources)