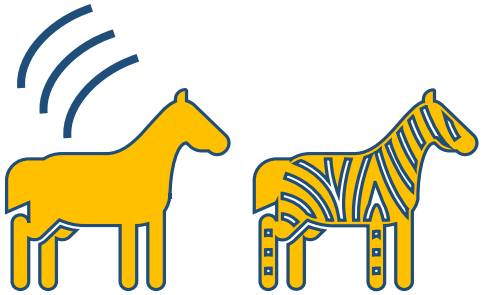
Zero-Shot Learning: The Vision-Language Wonder

Sandipan Sarma

Research Scholar

Dept. of CSE, IIT Guwahati



Technology Innovation Hub
IITG TIDF

The journey entails.....

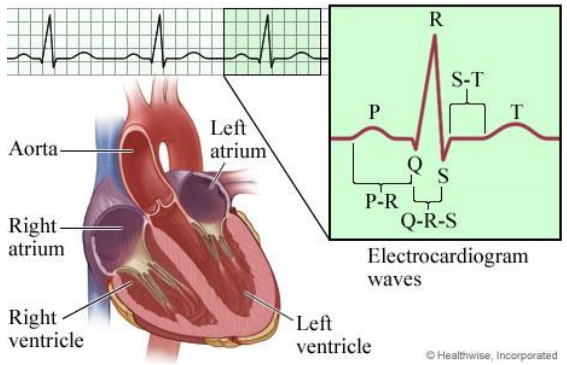
1. **Why** zero-shot learning?
2. **What** is zero-shot learning?
3. **How** to transfer knowledge in zero-shot learning?
4. **What** are the challenges in zero-shot learning?
5. **What** are the evaluation settings in zero-shot learning?
6. **Applications of** zero-shot learning
7. **Research at IITG** on zero-shot learning
8. **Q&A**

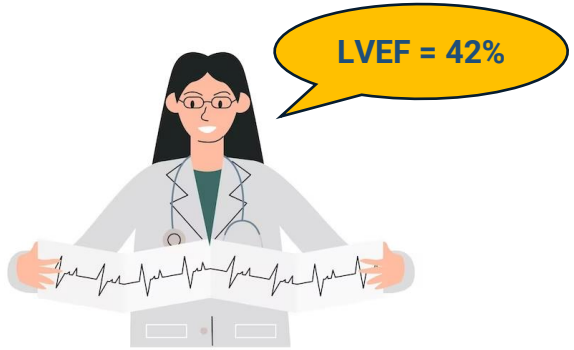
AI beats humans

Supervised learning

Task: Object recognition

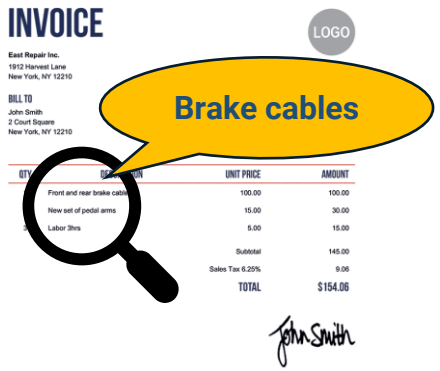
Human error: 5%
AI error: 2.2%





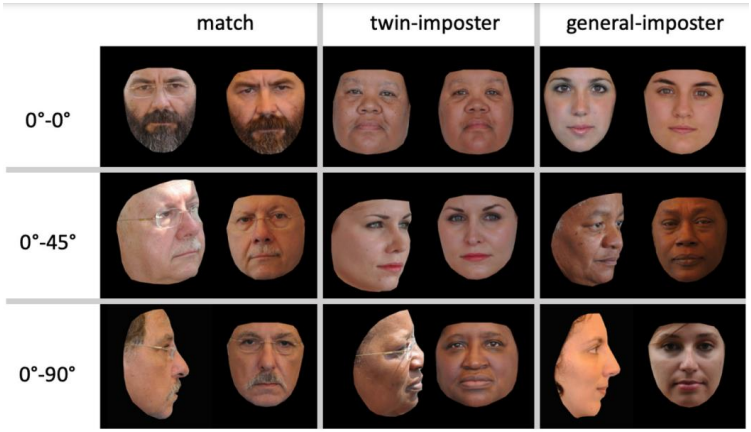
Task: Cardiac function assessment from ECG

Human error: 27.2%
AI error: 16.8%



Task: Structured document analysis

Human error: 10%
AI error: 2%



Task: Imposter detection

High correlation of human and AI predictions

Why ZSL?

- Supervised learning and its problems
- Possible solutions

Common learning paradigms

Why ZSL?


Killer Whale


Grizzly Bear


Gull

Supervised

Class?

Unsupervised

Similar patterns


Killer Whale

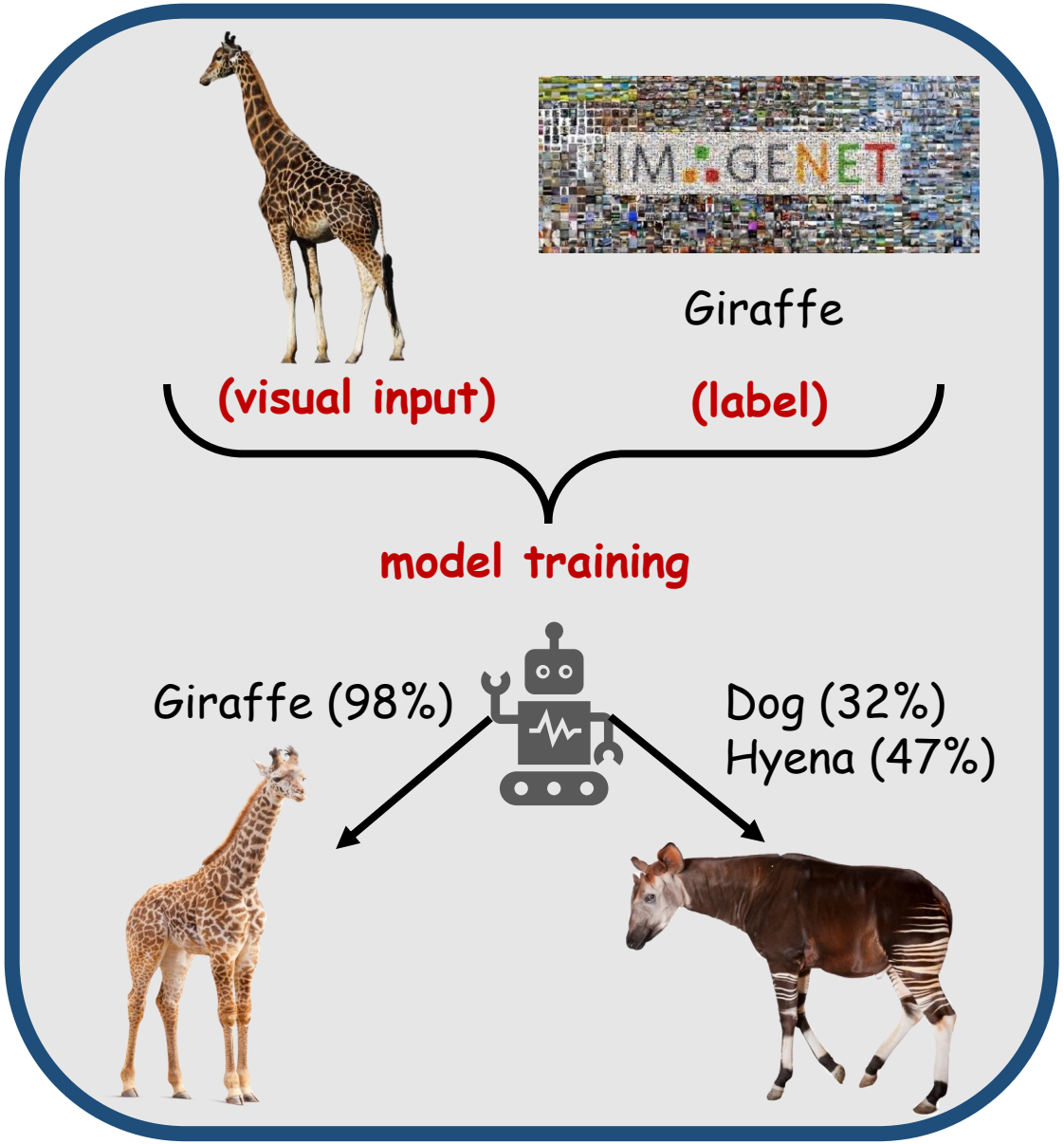

Grizzly Bear


Gull

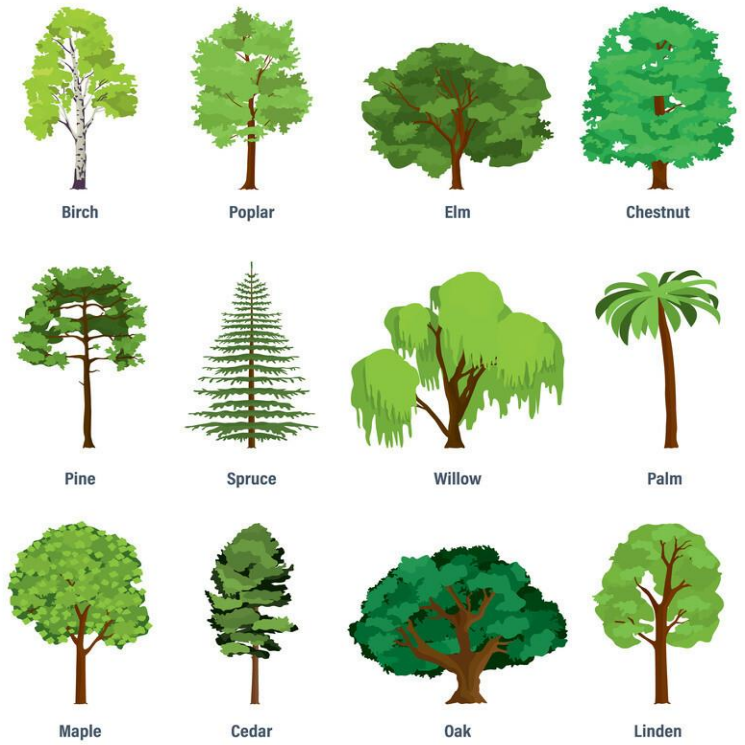
	black	white	blue	brown	gray	orange	red	yellow	patches	spots
antelope	-1.00	-1.00	-1.0	-1.00	12.34	0.0	0.0	0.0	16.11	9.19
grizzly+bear	39.25	1.39	0.0	74.14	3.75	0.0	0.0	0.0	1.25	0.00
killer+whale	83.40	64.79	0.0	0.00	1.25	0.0	0.0	0.0	68.49	32.69
beaver	19.38	0.00	0.0	87.81	7.50	0.0	0.0	0.0	0.00	7.50
dalmatian	69.58	73.33	0.0	6.39	0.00	0.0	0.0	0.0	37.08	100.00

Class?

Supervised learning and its problems



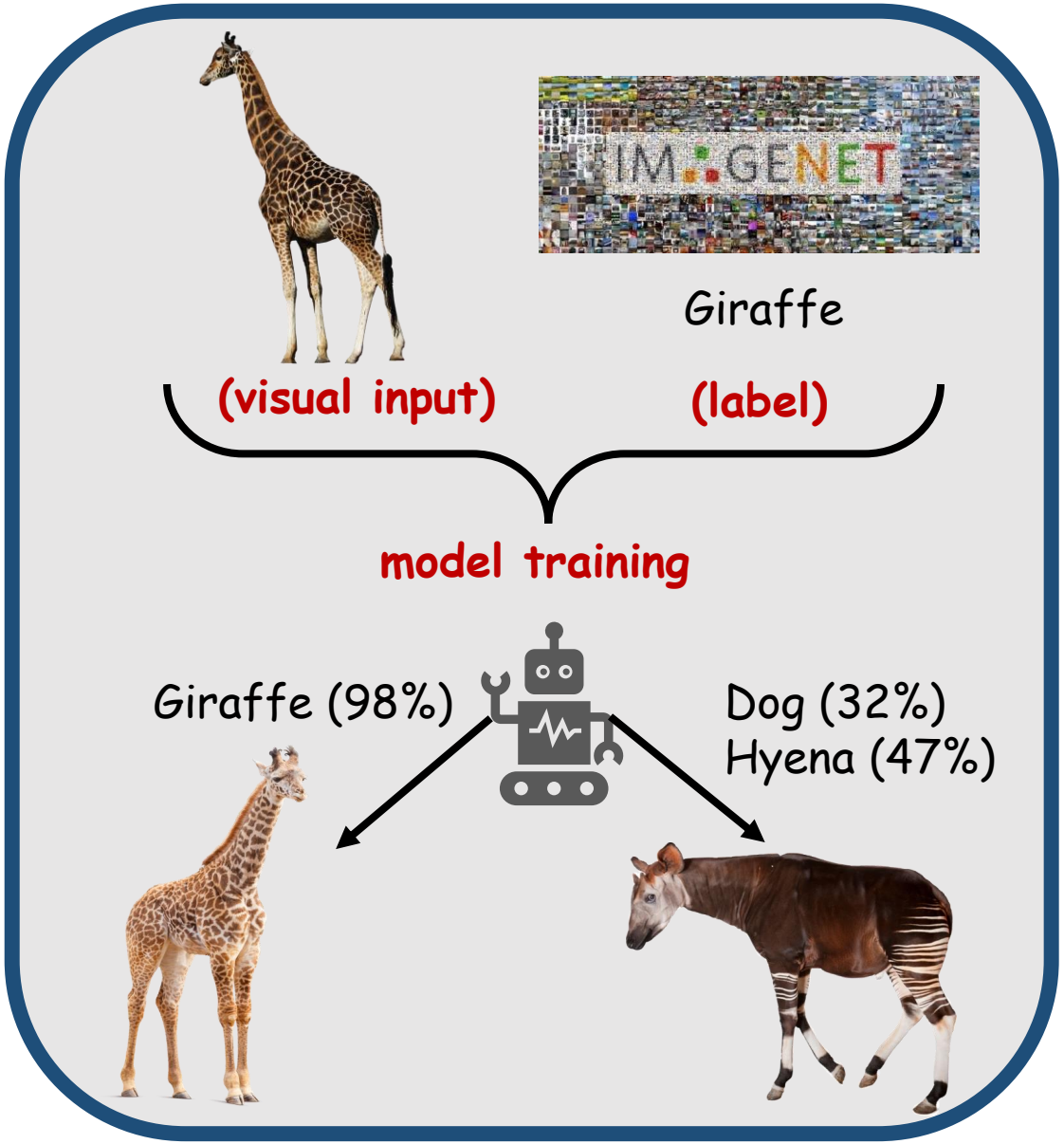
Why ZSL?



P1

Manual annotation needs **expert supervision** and is **time-consuming**

Supervised learning and its problems



General Class of Vehicles

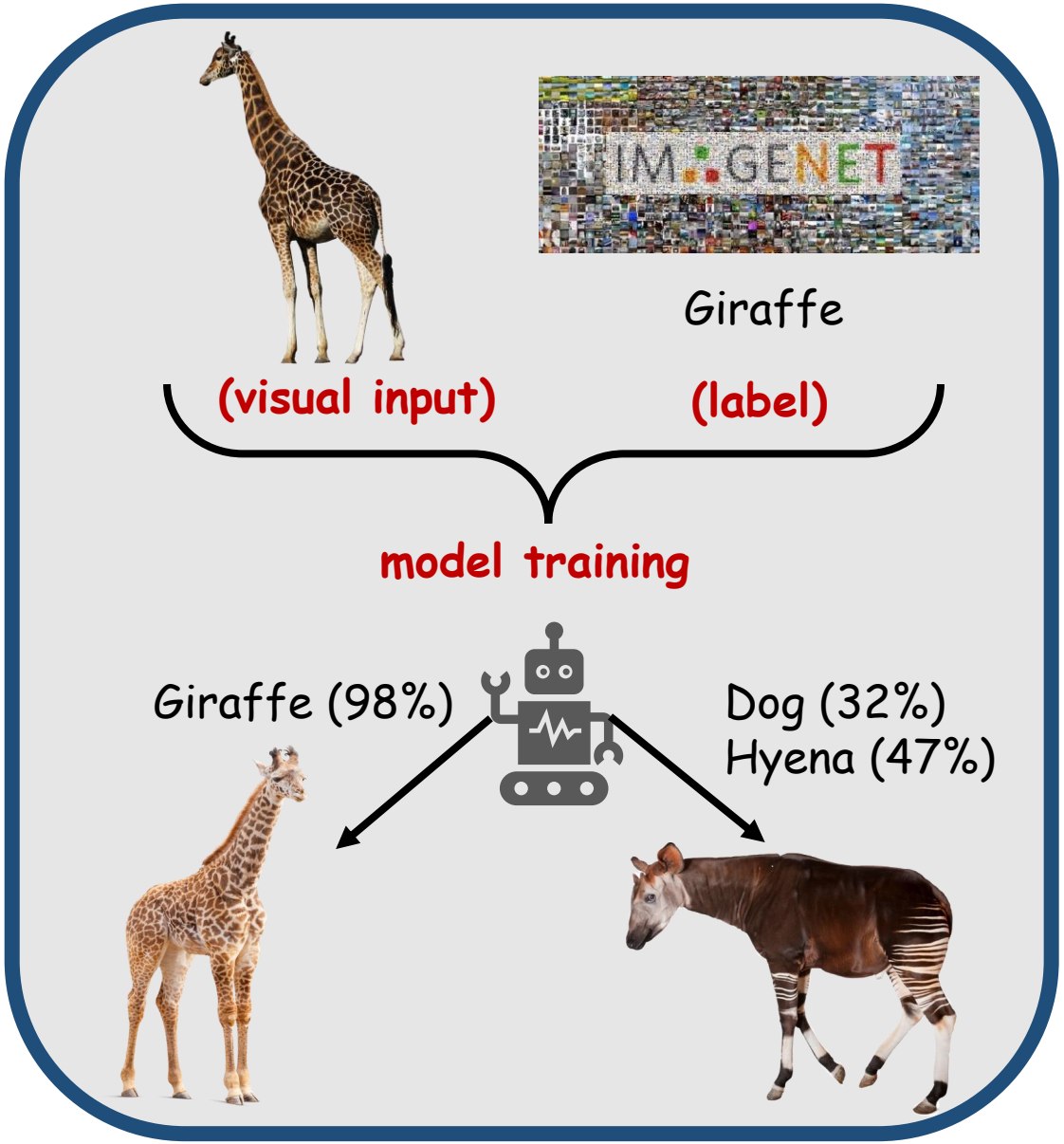
Annotated Sedan Cars (a sub-class of vehicles)

Unseen Concept Cars

P2

Changing **object appearances** over time

Supervised learning and its problems

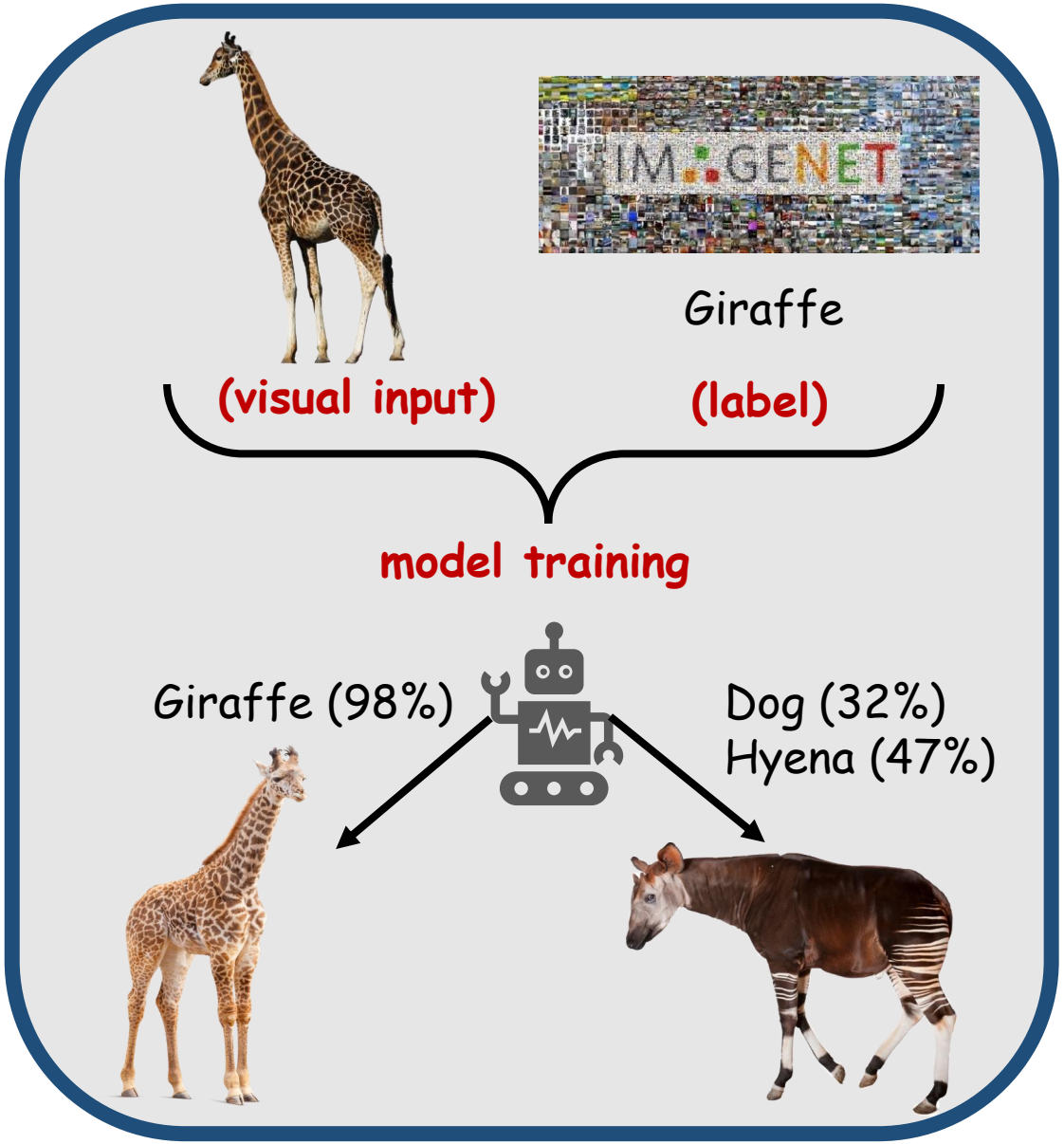


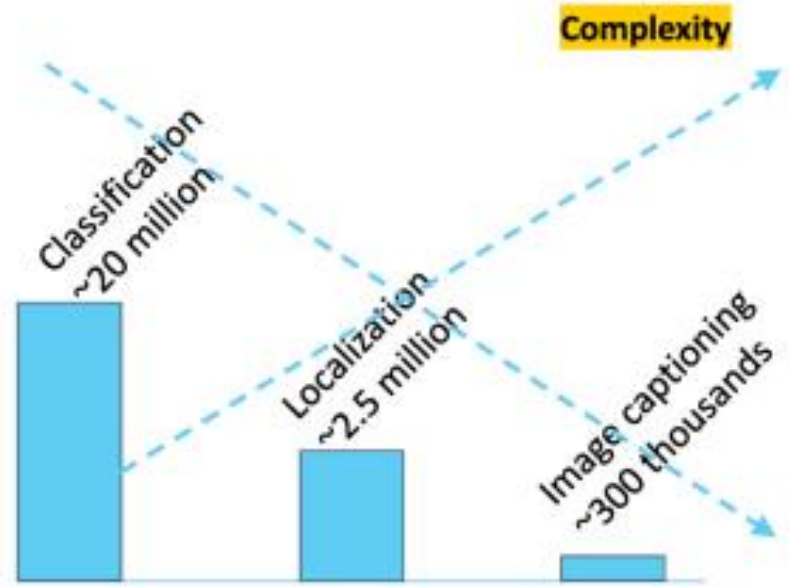
Platypus

Pangolin

P3 **Rare** objects in nature

Supervised learning and its problems





Imagenet+Open Images+MS COCO

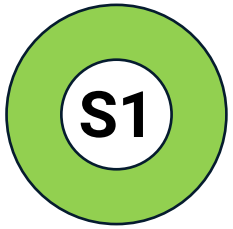
P4

Task complexity **increases,** size of annotated data **decreases**

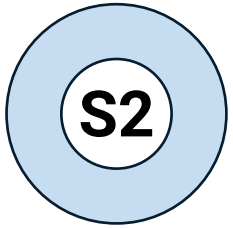
Possible solutions: N-Shot Learning

Shot : A single visual sample that a model can **see** during training

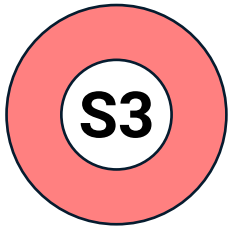
New solution : Learn with as few **shots** of an object as possible!



Few-shot Learning: Classify test image after training with **only a few images** (≤ 10) of that class!



One-shot Learning: Classify test image after training with **only one image** of that class!



Zero-shot Learning: Classify test image **without training on even a single image** of that class!

What is ZSL?

- Intuition
- Definition
- Visual embeddings
- Semantic embeddings

Intuition: Live ZSL demo

Audience task : Identify the ***Tarsier***

Never **seen** one before?

Is your human mind good at ZSL?

Additional information (object semantics)

Tarsier = Monkey-like body, furry, huge eyes, tree-climber

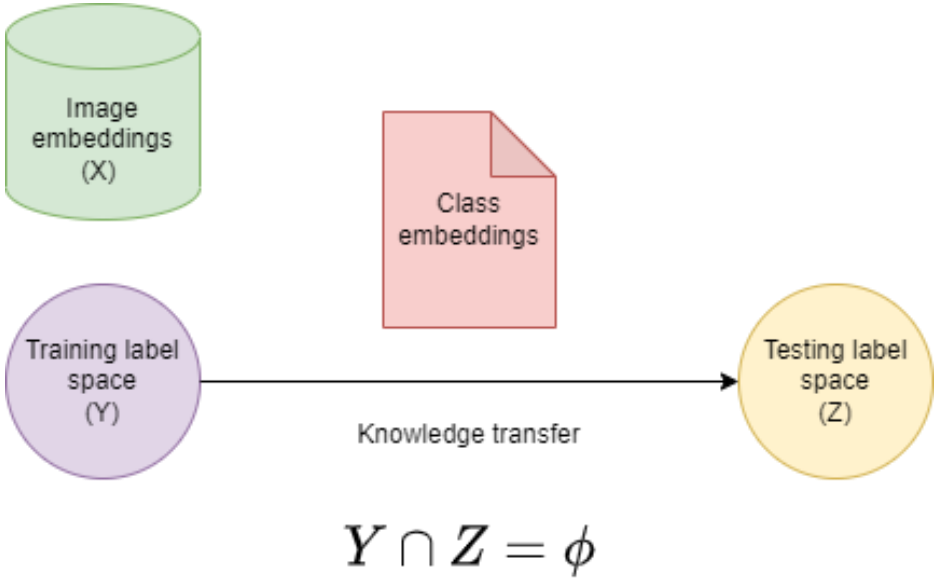
What's easy for the human mind, not so easy for the supervised models!

Zero-shot learning

Visual idea about ***seen*** class + **Semantic** idea = Visual – Semantic mapping → ***Unseen*** class recognition







$$Y \cap Z = \phi$$

	CLASSES	Image Embedding	Class Embedding
TRAINING	A	✔	✔
	B	✔	✔
	C	✔	✔
	D	✔	✔
	E	✔	✔
ZERO SHOT	F	✘	✔
	G	✘	✔
	H	✘	✔

How to obtain visual (image/video) embeddings?

AWA2 Dataset
(50 animals)

[News]: We introduce Proposed Splits Version 2.0 which fixed an issue in the original Proposed Split. We did not observe significant performance difference between the original Proposed Splits and Proposed Splits Version 2.0. More details can be found in this report. Please download Proposed Splits Version 2.0 below.

Paper

- CVPR17 paper on arxiv
- TPAMI paper on arxiv

Data Splits and Features for CUB, AWA1, AWA2, SUN and APY

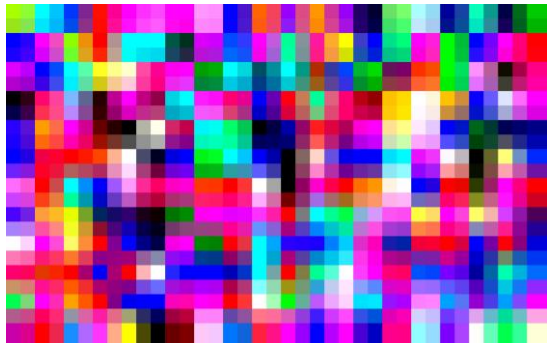
- Proposed Split Version 2.0
- Standard Split

Data Splits and Features for ImageNet

- ILSVRC2012 Res101 Feature
- ImageNet2011 Res101 Feature
- ImageNet Data Splits
- Scripts to read binary

AWA2 split [1]:

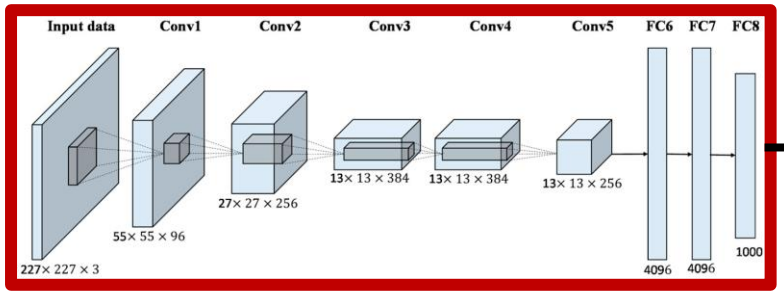
40 **seen** 10 **unseen**



Visual embeddings

Indian leopard

CNN feature extractor pretrained on



IMAGENET

[1] Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence, 41(9), 2251-2265.

How to obtain semantic/class embeddings?

MIT study [2], 1991

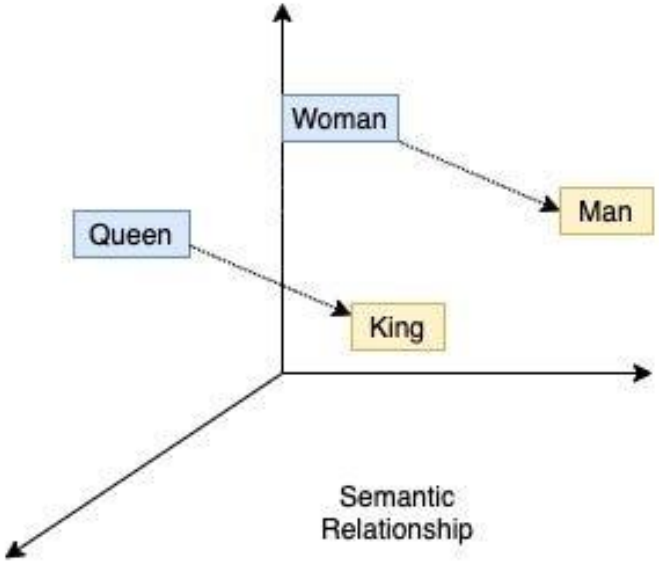
	black	white	blue	brown	gray	orange	red	yellow	patches	spots
antelope	-1.00	-1.00	-1.0	-1.00	12.34	0.0	0.0	0.0	16.11	9.19
grizzly+bear	39.25	1.39	0.0	74.14	3.75	0.0	0.0	0.0	1.25	0.00
killer+whale	83.40	64.79	0.0	0.00	1.25	0.0	0.0	0.0	68.49	32.69
beaver	19.38	0.00	0.0	87.81	7.50	0.0	0.0	0.0	0.00	7.50
dalmatian	69.58	73.33	0.0	6.39	0.00	0.0	0.0	0.0	37.08	100.00

Supervised way: Human-annotated

WIKIPEDIA
The Free Encyclopedia

word2vec

NeurIPS, 2013 [3]



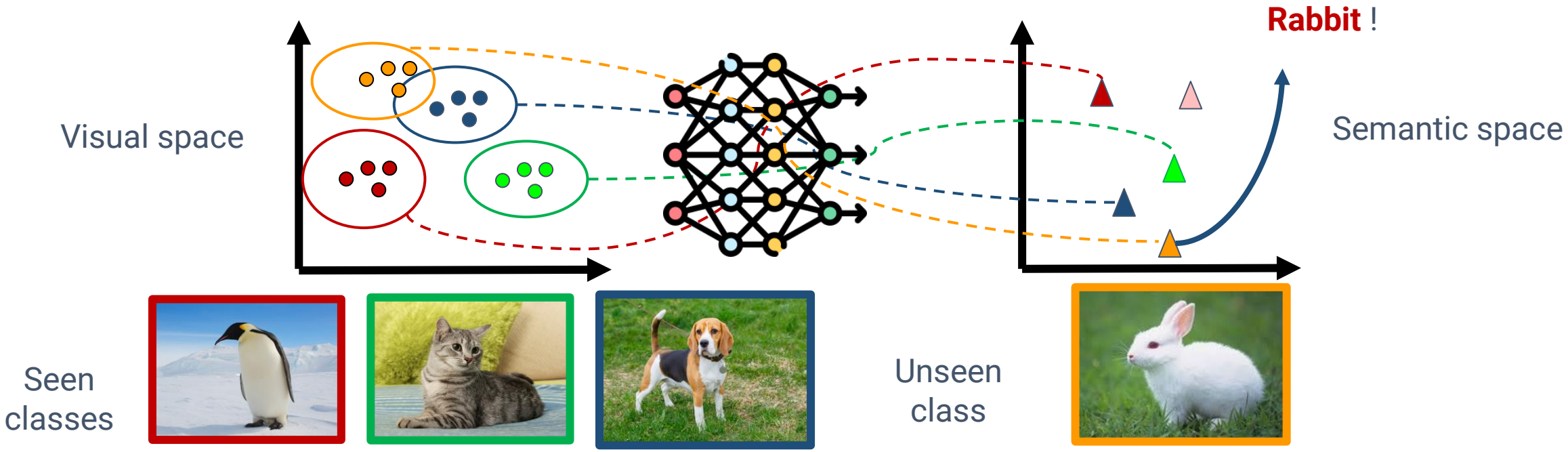
Unsupervised way: Pretrained language models

[2] Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. Cognitive Science, 15(2), 251-269.
[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

How to transfer knowledge in ZSL?

- Compatibility learning
- Visual data generation

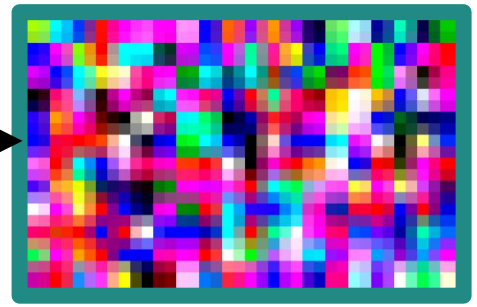
Method 1: Compatibility learning



$$F(x, y; W) = \theta(x)^T W \phi(y)$$

Method 2: Visual data generation

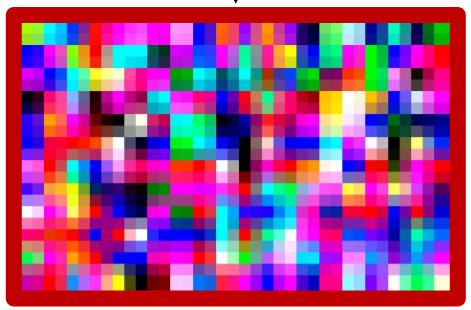
Indian leopard



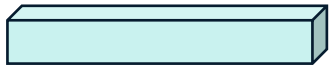
Visual embeddings (seen)

Semantic vector of unseen object (Word2Vec)

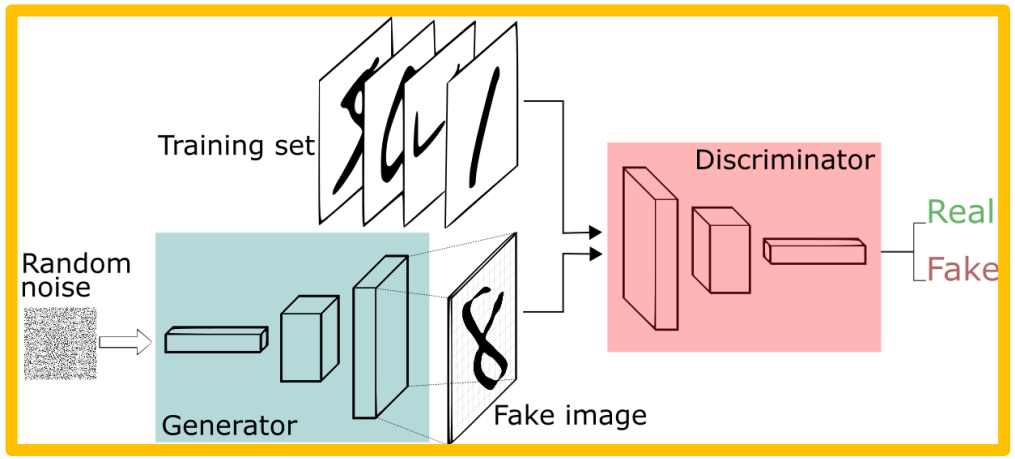
Trained GAN



Visual embeddings (unseen)



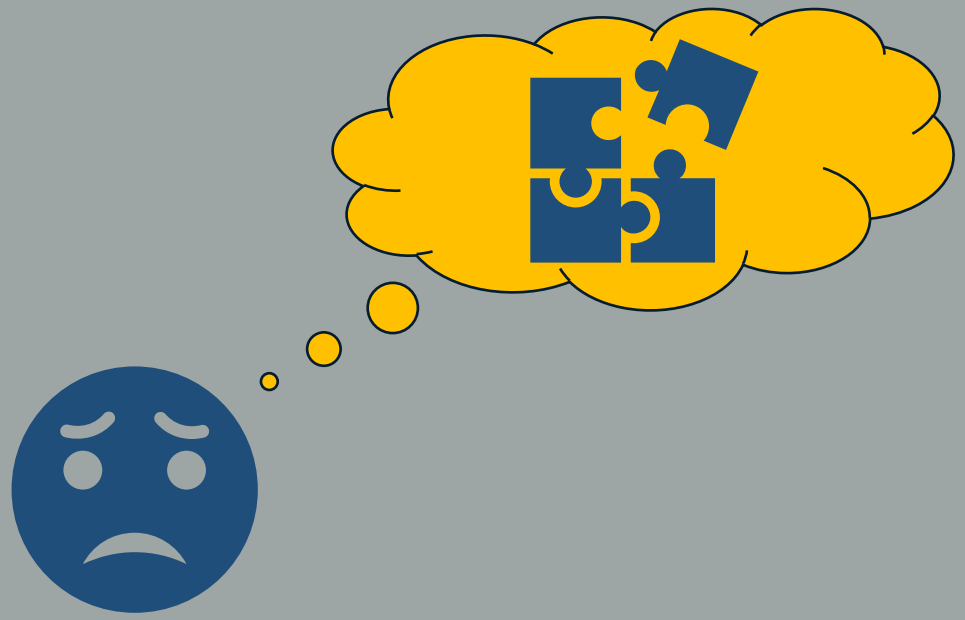
Semantic vector of seen object (Word2Vec)



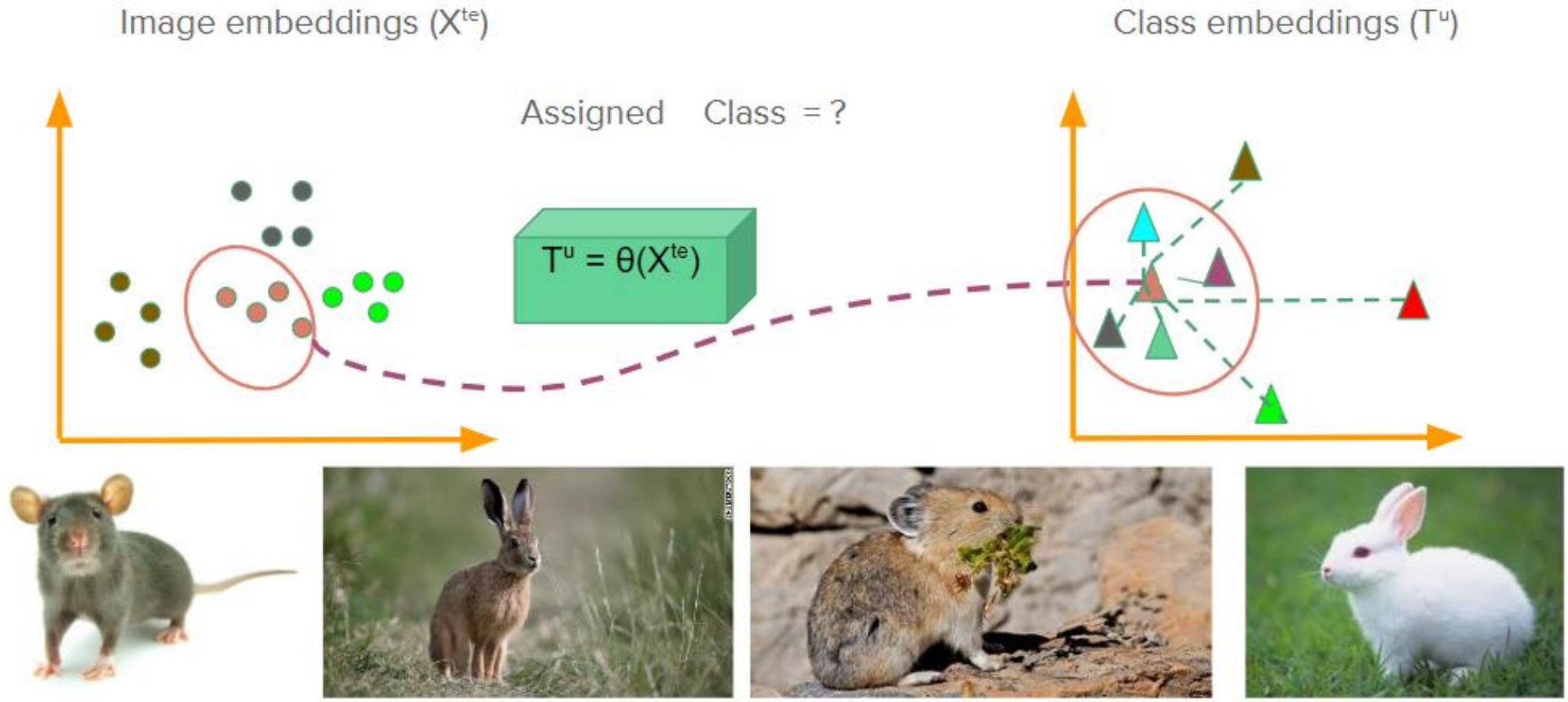
Train a Generative framework (e.g., GAN)

What are the challenges in ZSL?

- Hubness
- Domain shift
- Biasness



Hubness problem

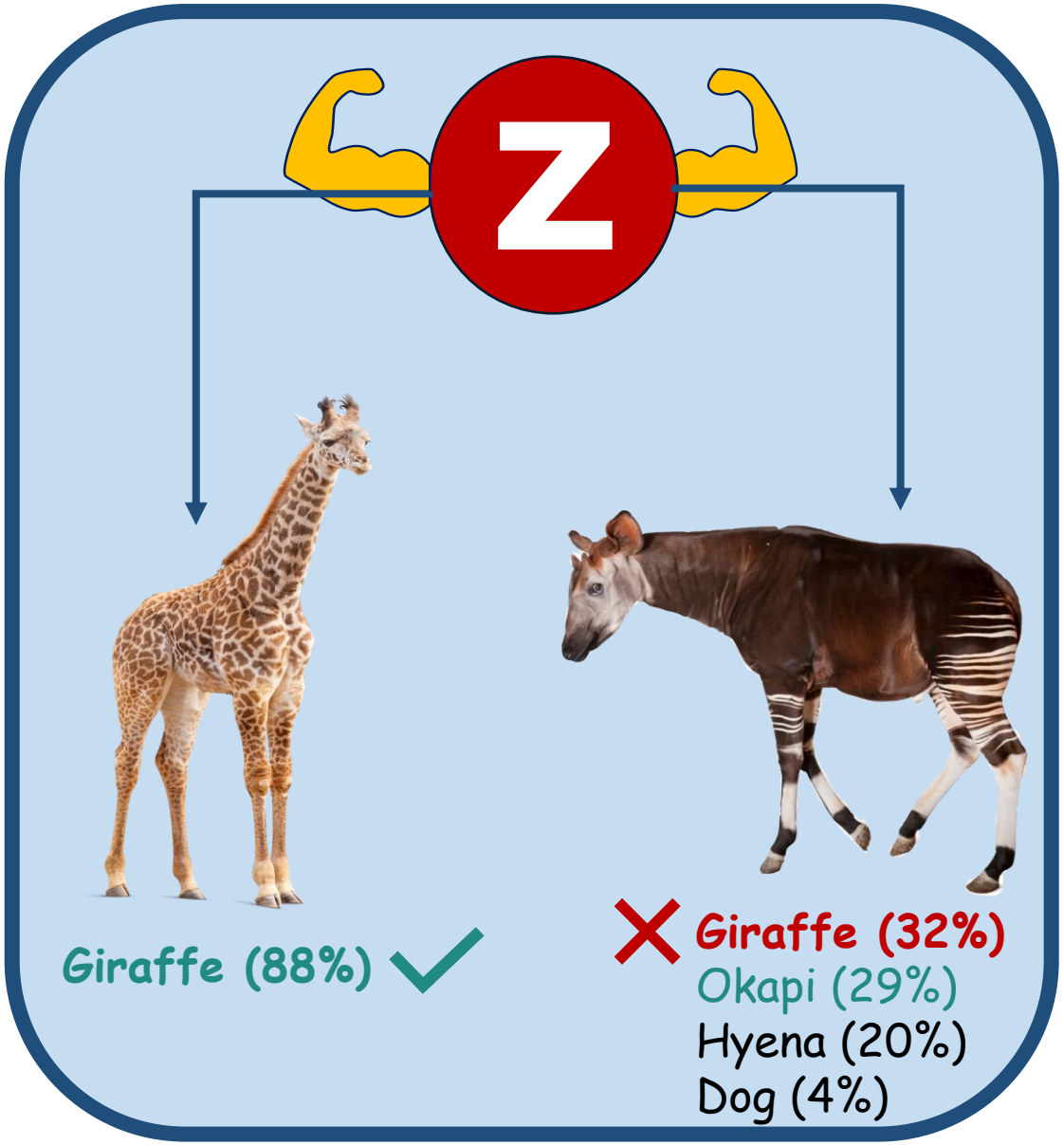


Semantic embedding of a certain class may become a **hub** for several similar classes

Domain shift problem

- **Every algorithm suffers** from domain shift – data distribution in training and testing data can be different
- **Extremely high in ZSL** since training and testing classes also disjoint
- In the given figure, due to difference in backgrounds, **even supervised models can misclassify** the bags!

Biasness problem

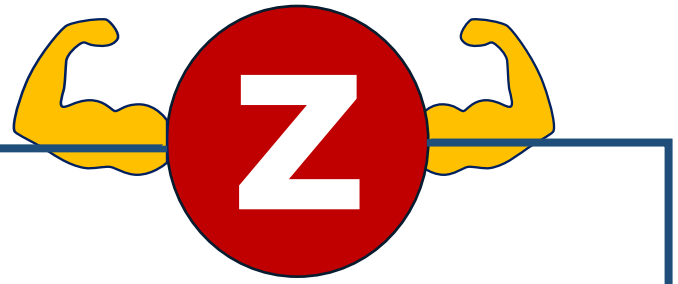


- While training, model **develops bias towards seen** class data
- At test time, semantically-similar unseen classes get **frequently misclassified** as one of the seen classes!

What are the evaluation settings in ZSL?

- Conventional and Generalized
- Multi-label

Conventional and Generalized ZSL



Trained ZSL model

Conventional setting

Test set has only unseen classes

Generalized setting (more practical)

Test set has both seen and unseen classes

Multi-label ZSL

Conventional ZSL prediction: beach

Multi-label ZSL prediction:
sand, beach, mountain, sky

- **Usual assumption**: An image belongs to a single class only
- **Multi-label setting**: Each image may belong to multiple classes

Applications of ZSL

Sandipan Sarma

Dept. of CSE, IITG

Tracking by natural language

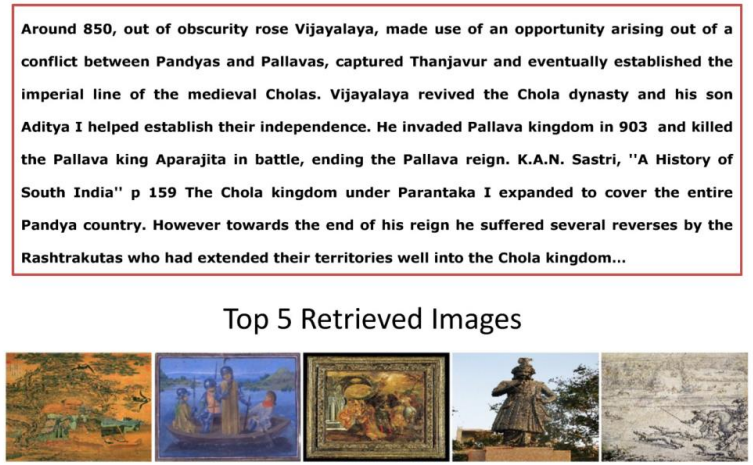
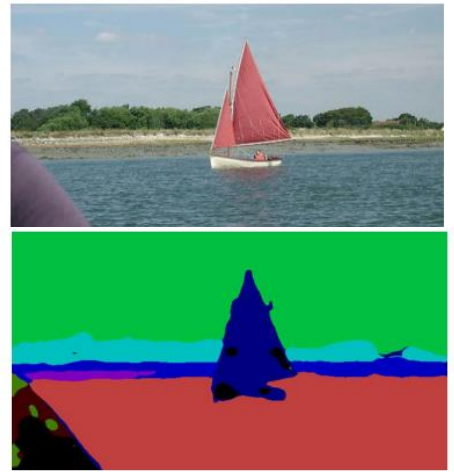
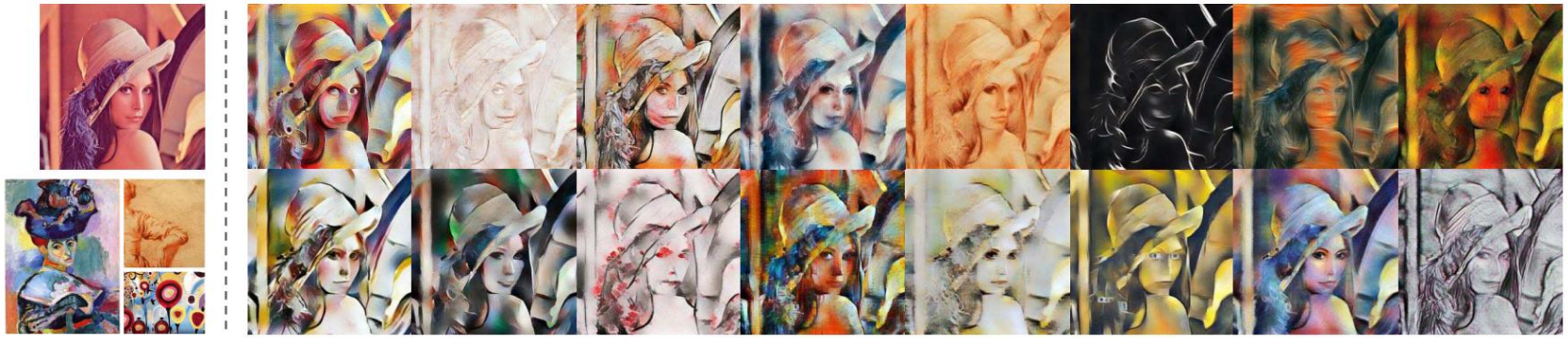


Image retrieval



Semantic segmentation



Style transfer

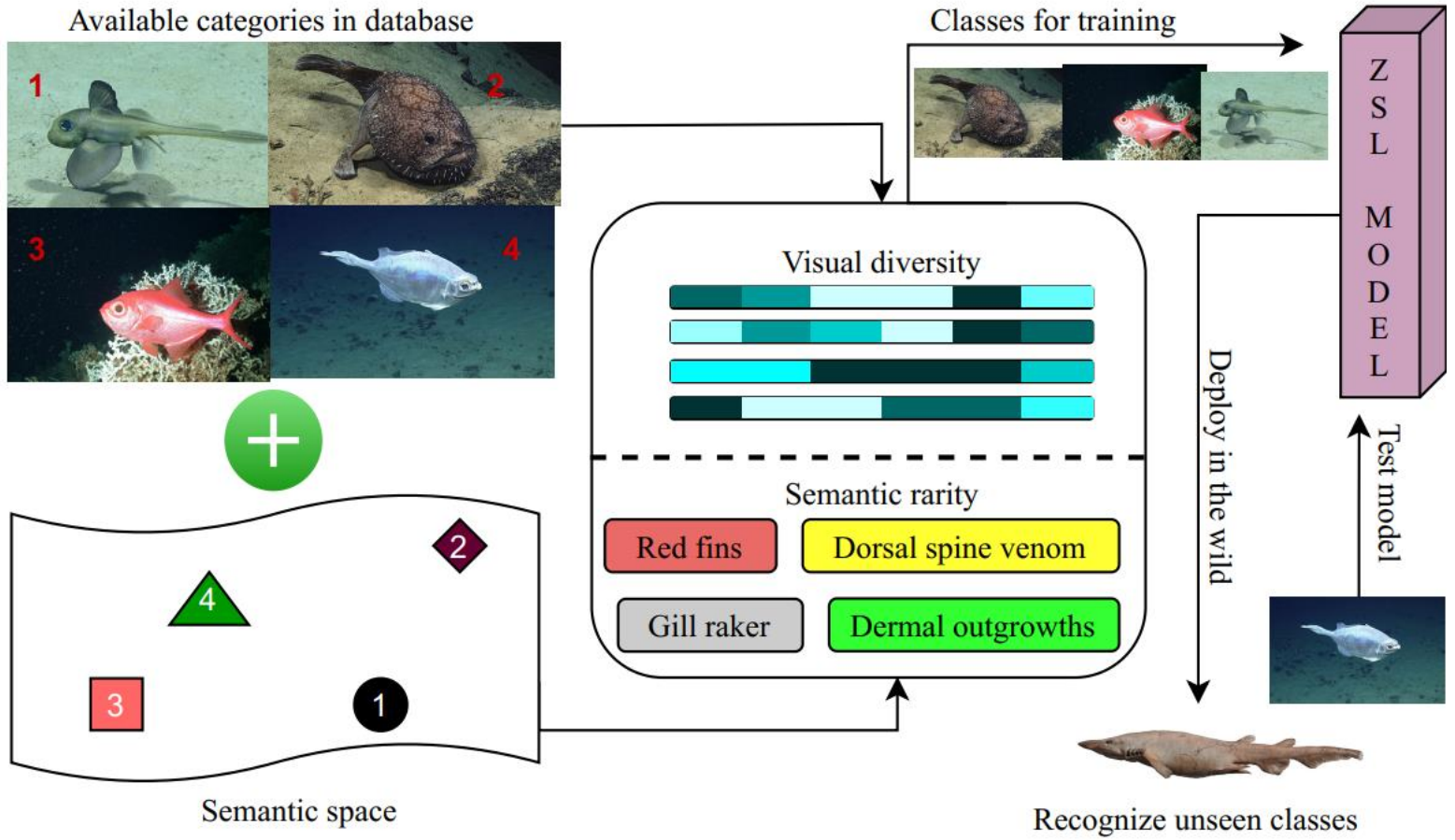
ZSL Research at CSE, IITG

- Object Recognition
- Object Detection
- Action Recognition
- Human-Object Interaction Detection
- Underwater Gesture Recognition

Sandipan Sarma

Dept. of CSE, IITG

Zero-shot object recognition

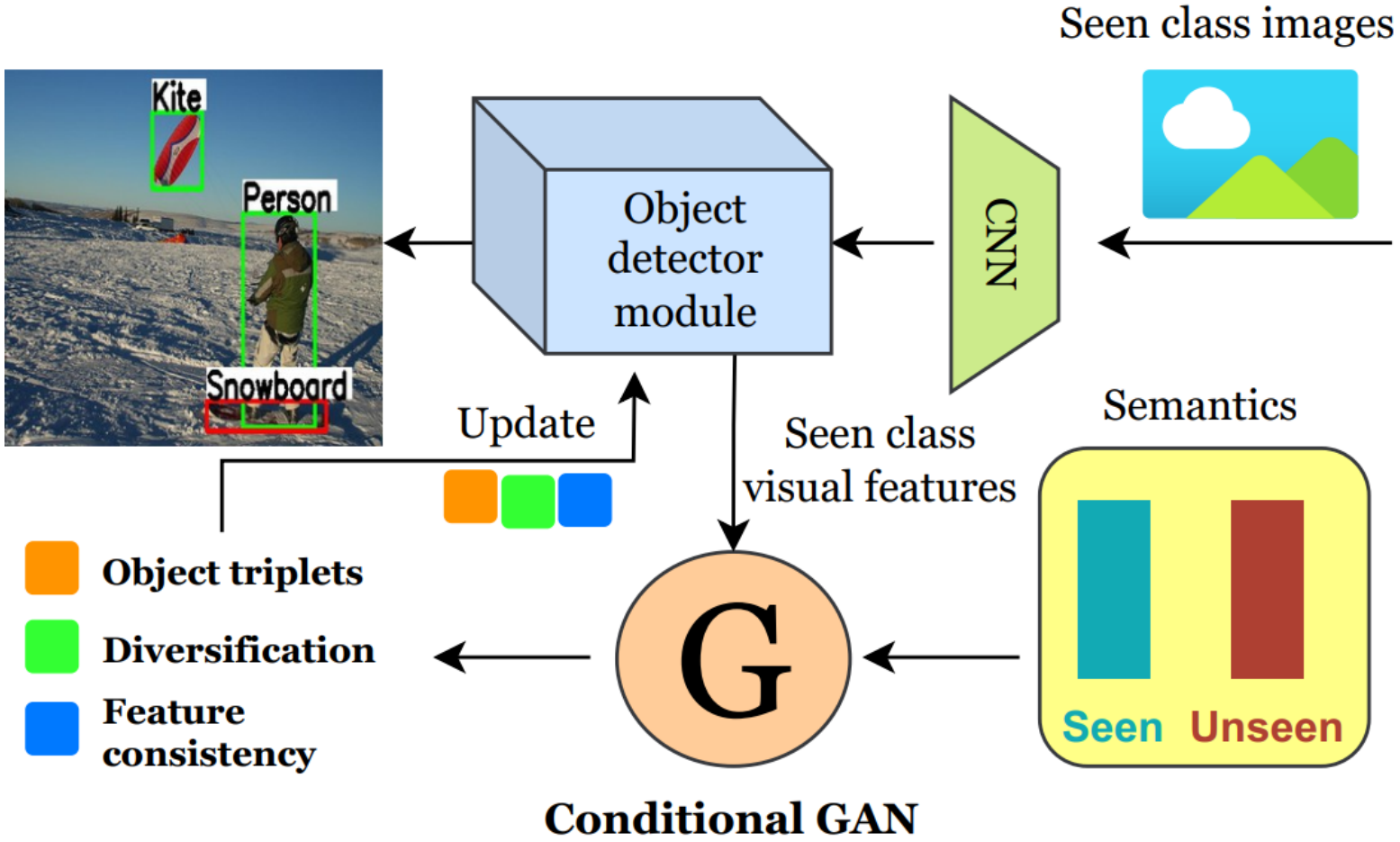


Explored the questions [4]:

1. Is it possible to **replace manual ZSL splits** with automatically-obtained splits?
2. Can **visual diversity and semantic rarity** in the object domain give a better idea to ZSL models?

[4] Sandipan Sarma and Arijit Sur. " ***DiRaC-I: Identifying Diverse and Rare Training Classes for Zero-Shot Learning***", in **ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM)**, vol. 20, no. 3, pp. 1-23, August 2023, doi: 10.1145/3603147.

Zero-shot object detection



Explored the questions [5]:

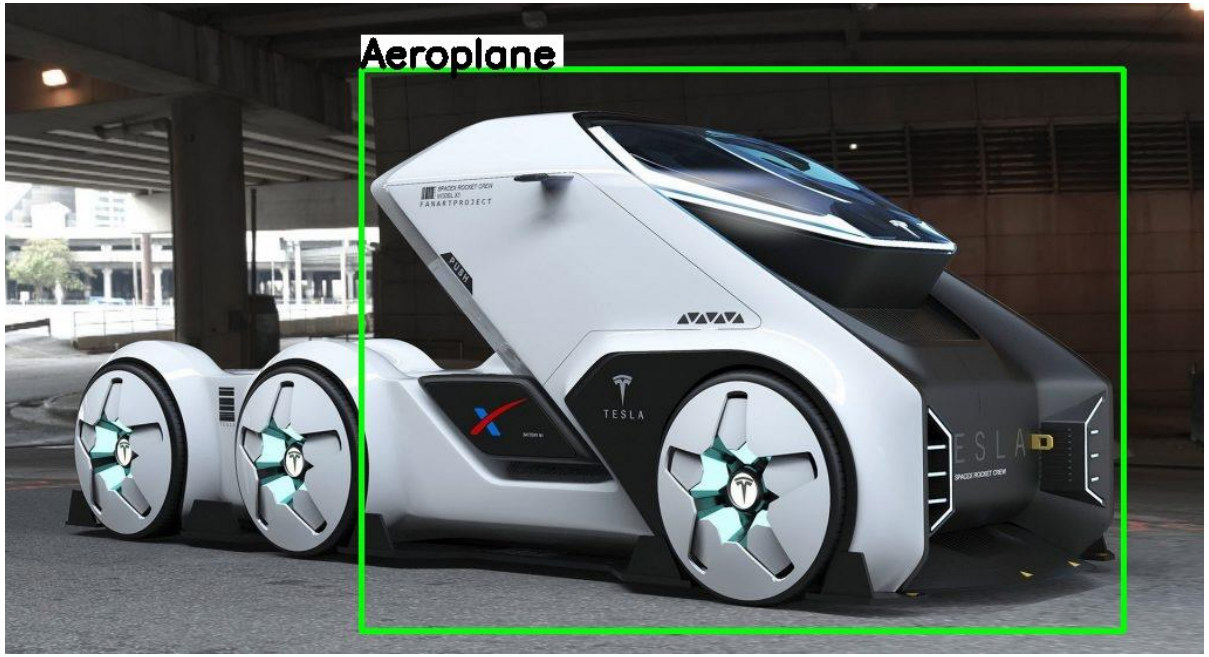
1. How to reduce **semantic confusion**?
2. How to maintain **visual-semantic consistency** during feature generation?

[5] Sandipan Sarma, Sushil Kumar, and Arijit Sur. 2022. ***Resolving Semantic Confusions for Improved Zero-Shot Detection*. In 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK**, November 21-24, 2022. BMVA Press. https://bmvc2022.mpi-inf.mpg.de/0347.pdf

In-the-wild data: Success cases

[5] Sandipan Sarma, Sushil Kumar, and Arijit Sur. 2022. ***Resolving Semantic Confusions for Improved Zero-Shot Detection*. In 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK**, November 21-24, 2022. BMVA Press. https://bmvc2022.mpi-inf.mpg.de/0347.pdf

In-the-wild data: Failure cases



[5] Sandipan Sarma, Sushil Kumar, and Arijit Sur. 2022. ***Resolving Semantic Confusions for Improved Zero-Shot Detection*. In 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK**, November 21-24, 2022. BMVA Press. https://bmvc2022.mpi-inf.mpg.de/0347.pdf

Motivation 1 : Object-centric duality

Hammer throw

Throw discus

Similar spatiotemporal motions, **distinguished** by objects

Boxing punching bag

Boxing speed bag

Similar interactions with **functionally-similar** objects

Motivation 2 : Environment-centric duality

Breaststroke

Horse riding

Action environment can add **distinguishability**

Rafting

Kayaking

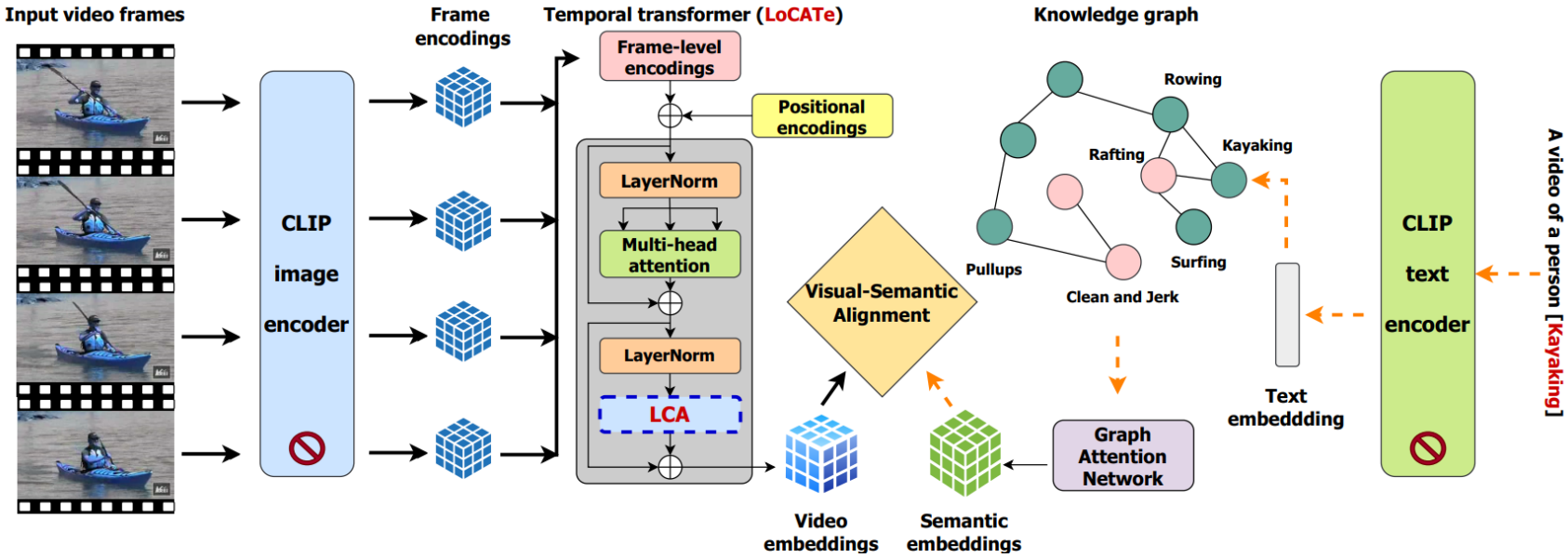
Similar interactions in specific **environments**

Zero-shot action recognition

UCF

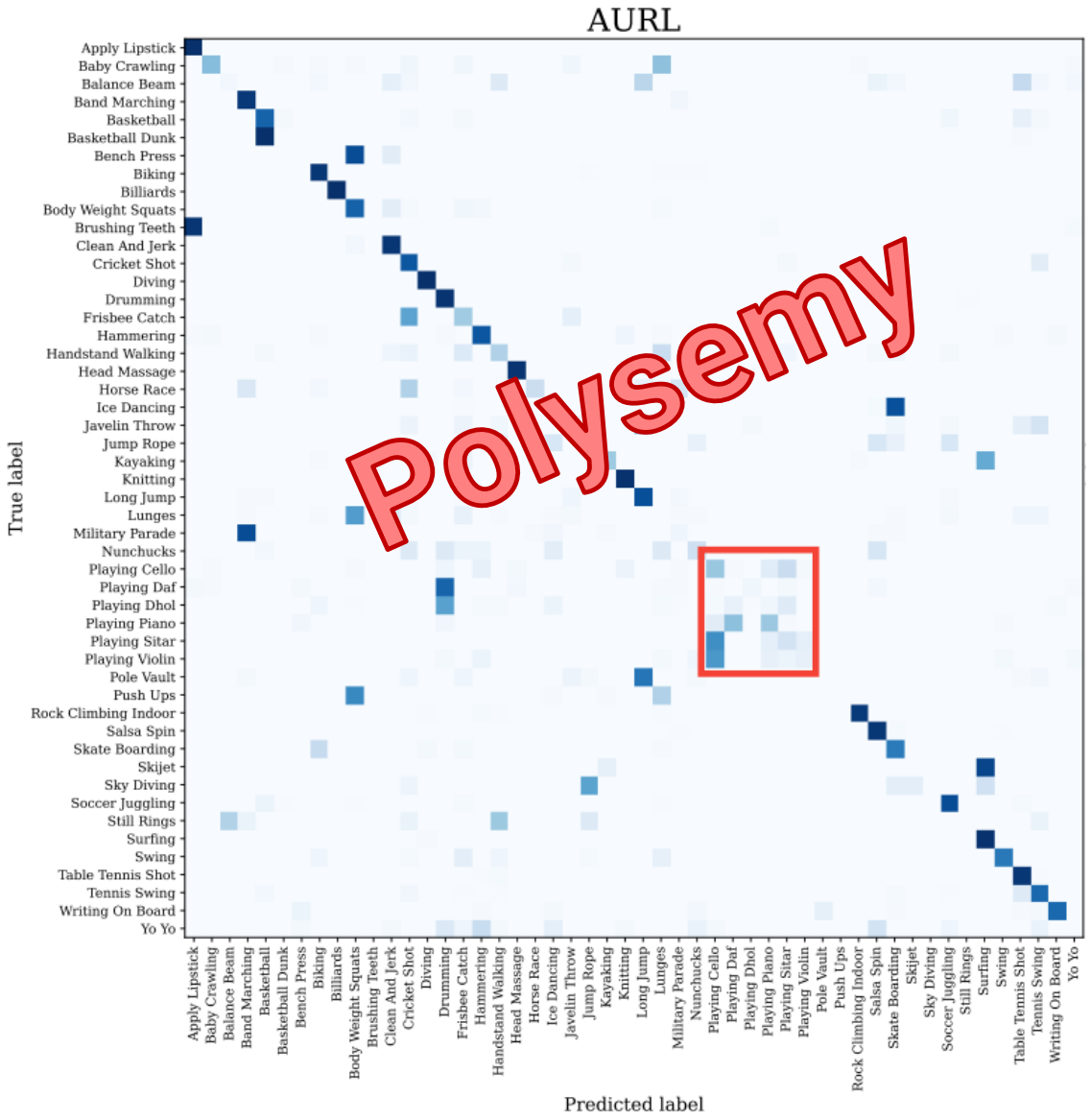
Training

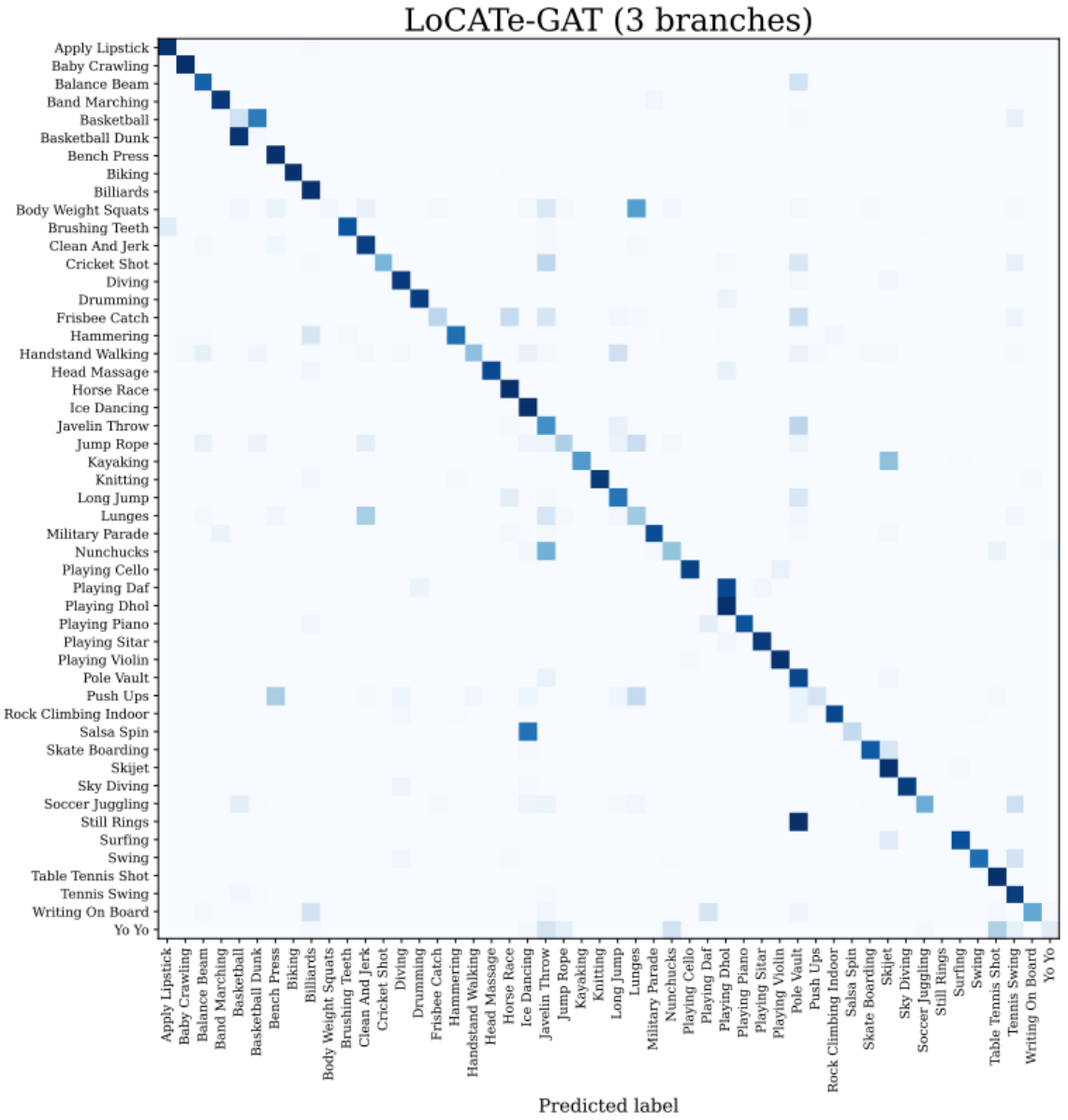
ZSL model



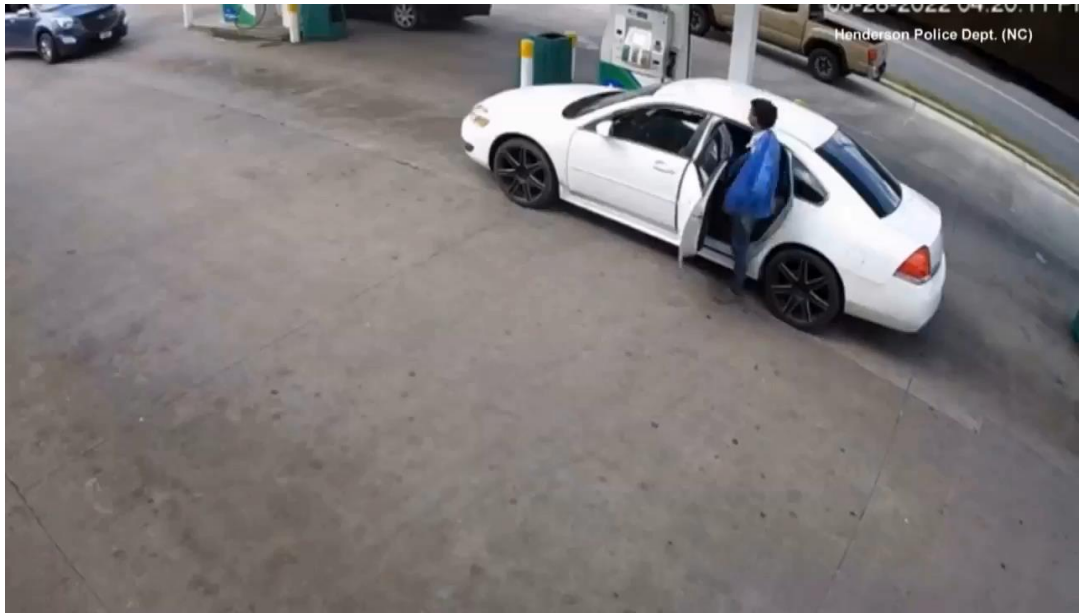
Training on regular activities: Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Shooting, Basketball Dunk, Bench Press, Biking, Javelin Throw, Playing Piano,……

Zero-shot action recognition





Zero-shot action recognition



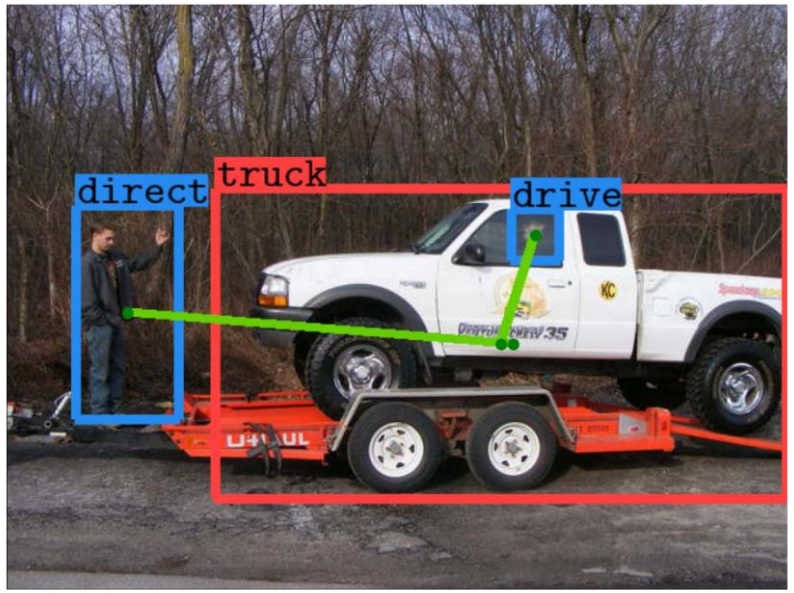
Testing on suspicious activities:
Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, Vandalism

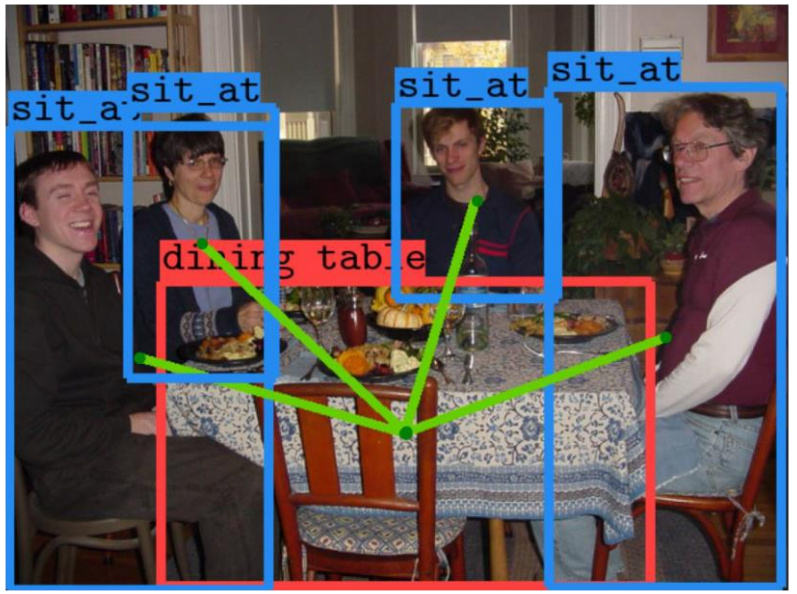
[6] Sandipan Sarma, Divyam Singal, and Arijit Sur. "***LoCATe-GAT: Modeling Multi-Scale Local Context and Action Relationships for Zero-Shot Action Recognition***", in **IEEE Transactions on Emerging Topics in Computational Intelligence (IEEE TETCI)**, November 2024, doi:10.1109/TETCI.2024.3499995.

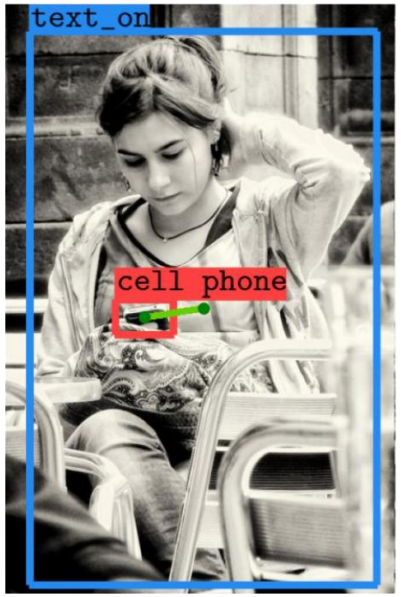
Zero-shot human-object interaction detection

Aim: For an input image, output a set of bounding box **interactive** pairs, each localizes a human plus an object and predicts an HOI class label

Data labels: <human, relationship, object> triplets. Different HOIs may share the same human, action or object.







[7] Sandipan Sarma, Pradnesh Kalkar, and Arijit Sur. 2024. ***Boosting Zero-shot Human-Object Interaction Detection with Vision-Language Transfer***. In 49th **IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 14-19 April, 2024, Seoul, South Korea**

Why Zero-Shot UGR?

Example gestures from each of the 16 classes in the **CADDY dataset** **[4]**

Real-time deployment

Zero-shot underwater gesture recognition

Since there are no prior works in ZSUGR, we adapted a few *zero-shot image classification methods* for the ZSUGR task, following previous works on zero-shot gesture recognition [8, 9]

Pretrained CNN	Supervised	S_{gzsl}	U_{gzsl}	H
AlexNet [62]	82.89	71.75 ± 2.19	0.76 ± 1.21	1.48 ± 2.35
VGG-16 [62]	95.00	76.20 ± 1.23	0.42 ± 0.57	0.84 ± 1.12
ResNet-18 [66]	98.00	53.54 ± 3.77	0.72 ± 1.25	1.38 ± 2.39
ResNet-50 [63]	97.06	61.94 ± 2.86	0.79 ± 1.27	1.53 ± 2.44
GoogleNet [62]	90.08	53.19 ± 3.46	0.8 ± 1.39	1.53 ± 2.64
MobileNet-v3 [66]	84.32	62.78 ± 5.21	1.31 ± 2.23	2.45 ± 4.18

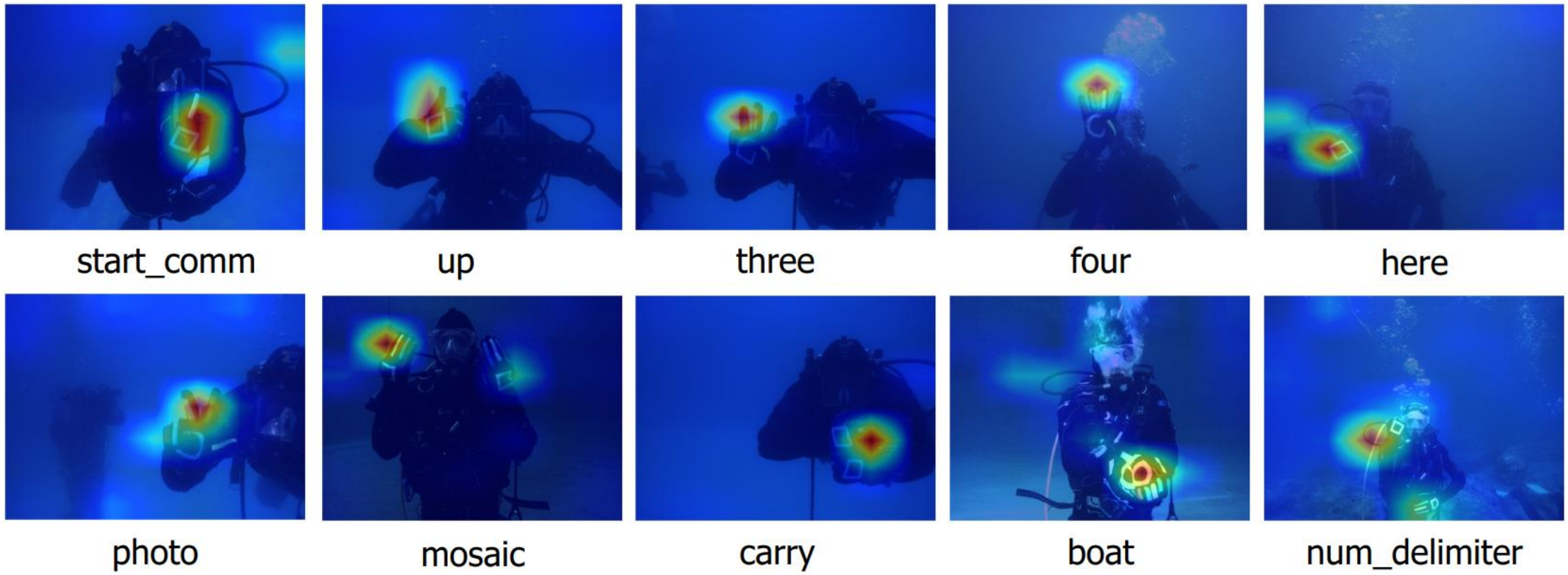
Top-1 accuracy (in %) achieved by **supervised** pretrained CNN models

Method	U_{czsl}	S_{gzsl}	U_{gzsl}	H
TFVAEGAN [103]	41.50 ± 5.47	79.57 ± 13.11	13.49 ± 5.28	22.51 ± 7.40
CNZSL [104]	16.72 ± 0.08	20.27 ± 3.84	11.88 ± 2.78	14.97 ± 3.23
FREE [105]	15.83 ± 3.95	84.88 ± 9.03	14.55 ± 3.36	24.61 ± 4.50
CE-GZSL [106]	39.89 ± 6.49	$\mathbf{94.11 \pm 0.55}$	2.58 ± 0.45	5.01 ± 0.87
DGZ [13]	-	57.89 ± 2.90	15.72 ± 3.32	24.62 ± 4.14
Ours	$\mathbf{45.91 \pm 4.71}$	61.93 ± 5.71	$\mathbf{20.03 \pm 7.14}$	$\mathbf{29.53 \pm 7.06}$

Top-1 accuracy (in %) achieved by **our zero-shot model**

[8] N. Madapana, "Zero-shot learning for gesture recognition," in Proceedings of the 2020 international conference on multimodal interaction, 2020, pp. 754–757.
[9] J. Wu, Y. Zhang, and X. Zhao, "A prototype-based generalized zero-shot learning framework for hand gesture recognition," in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 3435–3442.

Zero-shot underwater gesture recognition



[10] Sandipan Sarma, Gundameedi Sai Ram Mohan, Hariansh Sehgal, and Arijit Sur. "***Zero-Shot Underwater Gesture Recognition***", **27th International Conference on Pattern Recognition (ICPR) 2024, Kolkata, India**, pp. 346–361, doi: 10.1007/978-3-031-78183-4 22

Thank you !
Questions?

sandipan.sarma@iitg.ac.in

github.com/sandipan211

linkedin.com/in/sandipan-sarma

Technology
Innovation Hub
IITG TIDF