# Rethinking Generalization in Vision Models: Architectures, Modalities, and Beyond

Ziwei Liu

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

# Why Need Generalization?

- In practice there is often a distribution shift between training and testing

# Rethinking Generalization

***Corruptions / Perturbations / Domain Shifts***

**Covariate Shift**

# Rethinking Generalization

**Semantic Shift**

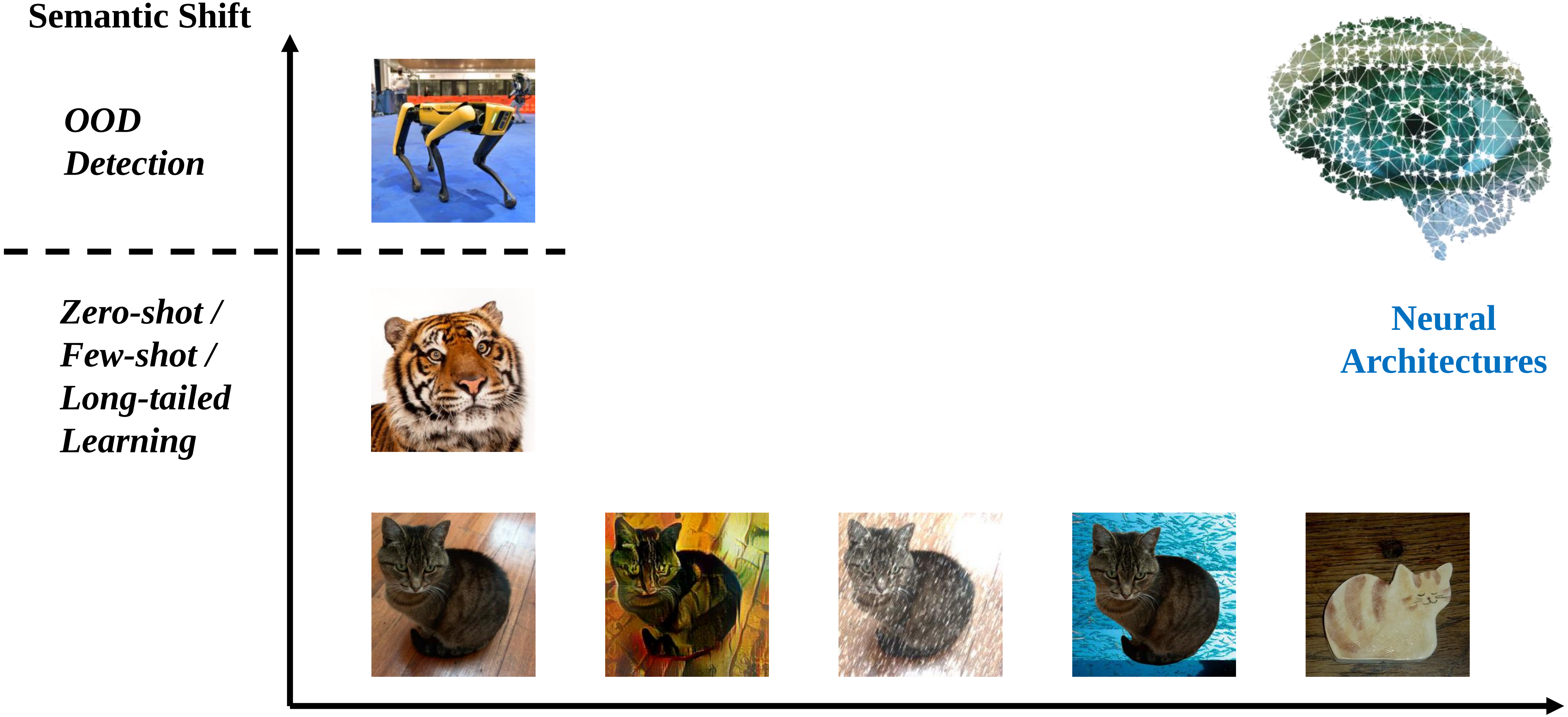
***OOD Detection***

***Zero-shot / Few-shot / Long-tailed Learning***

***Corruptions / Perturbations / Domain Shifts***

**Covariate Shift**

# Generalization in Vision Models



**Semantic Shift**

***OOD Detection***

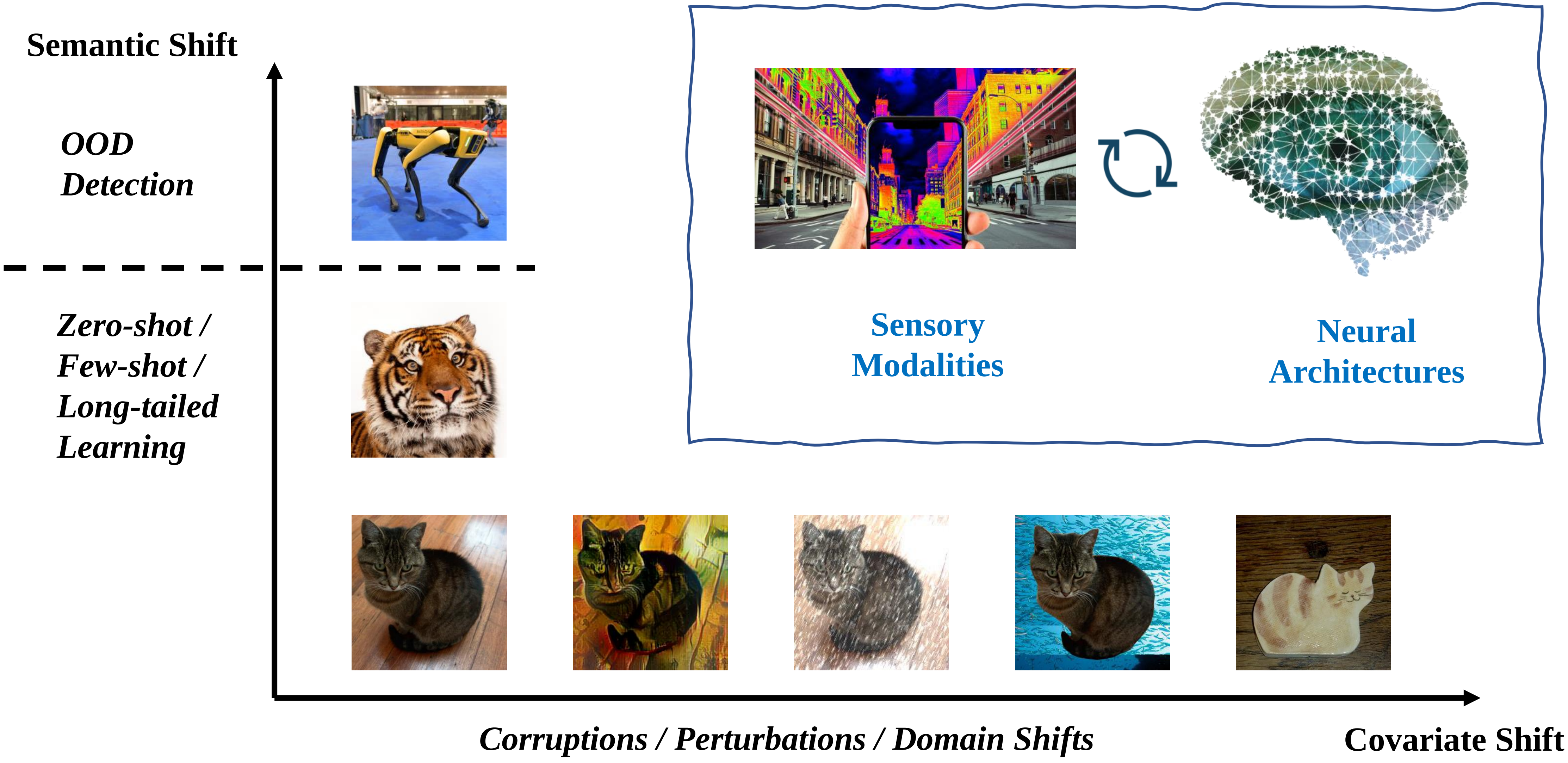
***Zero-shot / Few-shot / Long-tailed Learning***

**Neural Architectures**

***Corruptions / Perturbations / Domain Shifts***

**Covariate Shift**

# Generalization in Vision Models



**Semantic Shift**

***OOD Detection***

***Zero-shot / Few-shot / Long-tailed Learning***

**Sensory Modalities**

**Neural Architectures**

***Corruptions / Perturbations / Domain Shifts***

**Covariate Shift**

# Generalization in Vision Models

Semantic Shift

*OOD Detection*

*Zero-shot / Few-shot / Long-tailed Learning*

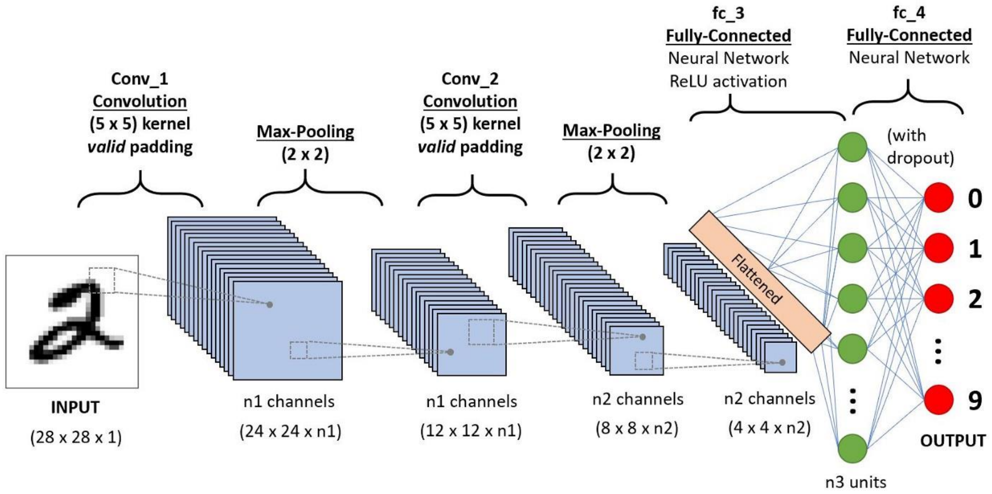
Sensory Modalities

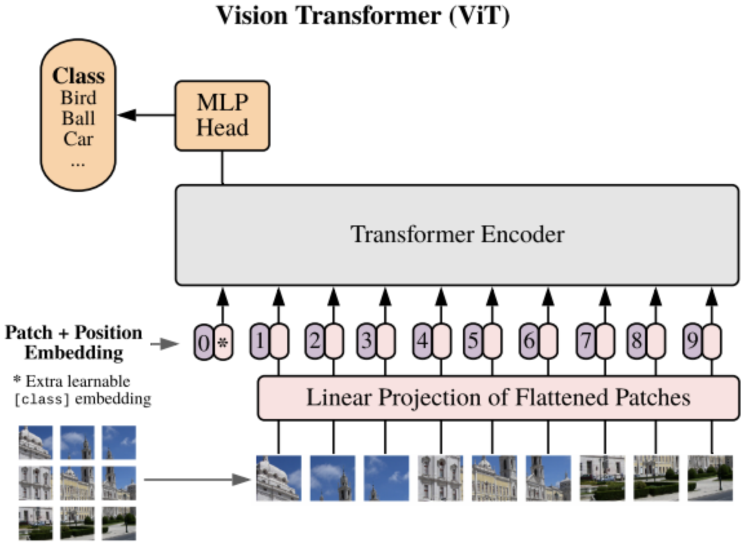
**Neural Architectures**

***Corruptions / Perturbations / Domain Shifts***

**Covariate Shift**

# Convolution v.s. Attention (2D Vision)





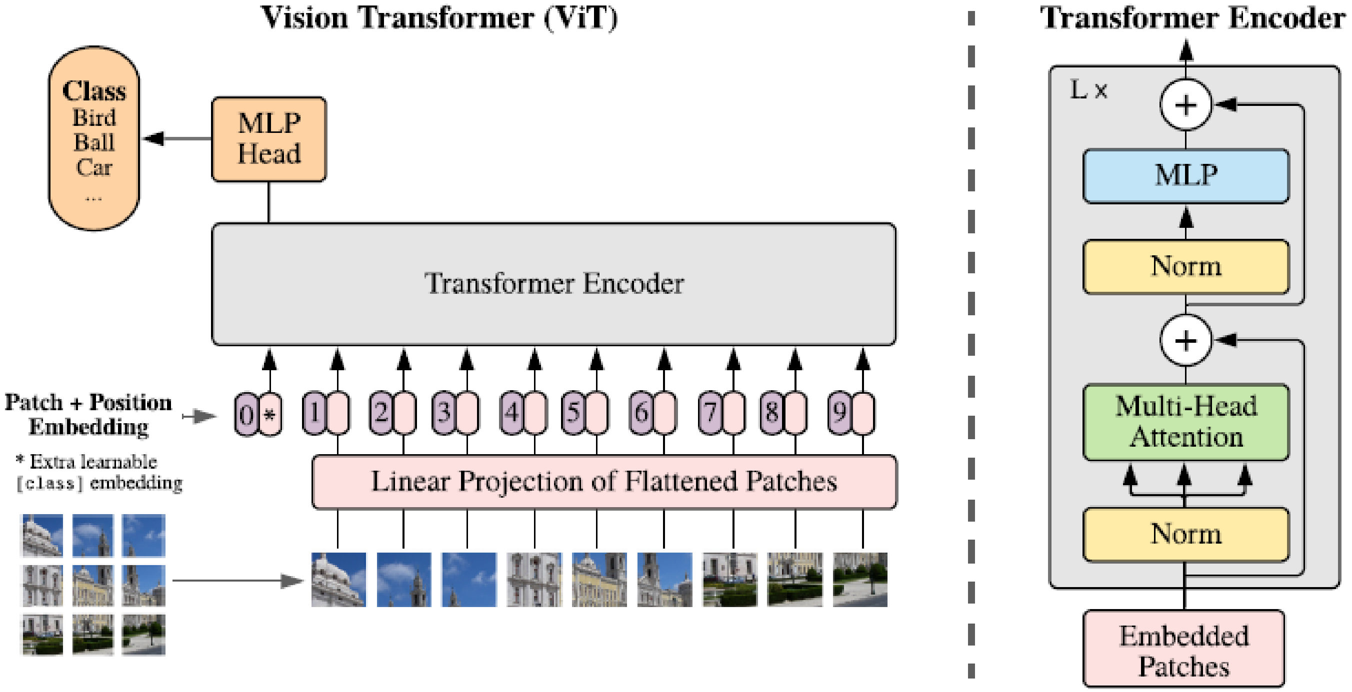
Zhang et al., Delving Deep into the Generalization of Vision Transformers under Distribution Shifts, CVPR 2022

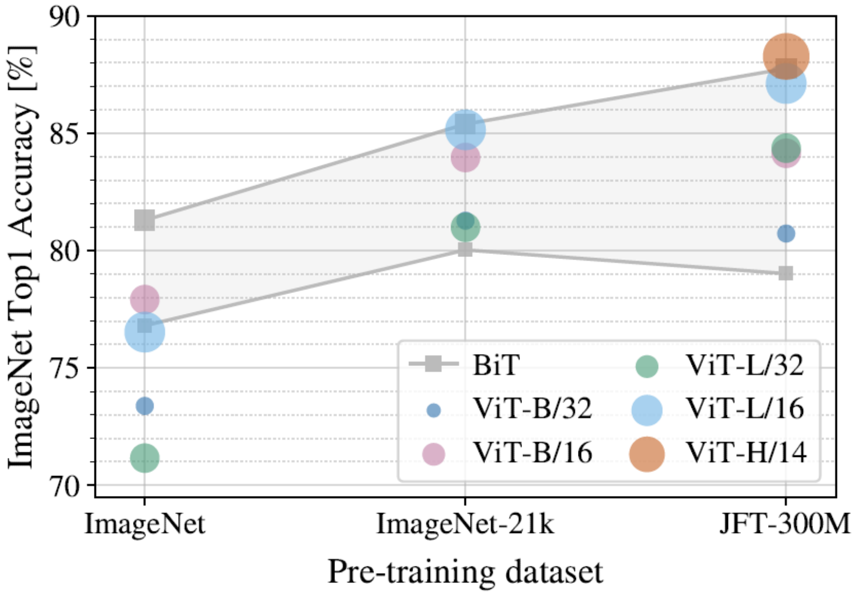
Related Works:

Bai et al., Are Transformers More Robust Than CNNs, NeurIPS 2021
Zhou et al., Understanding the Robustness in Vision Transformers, arXiv 2022

# The Rise of Transformers

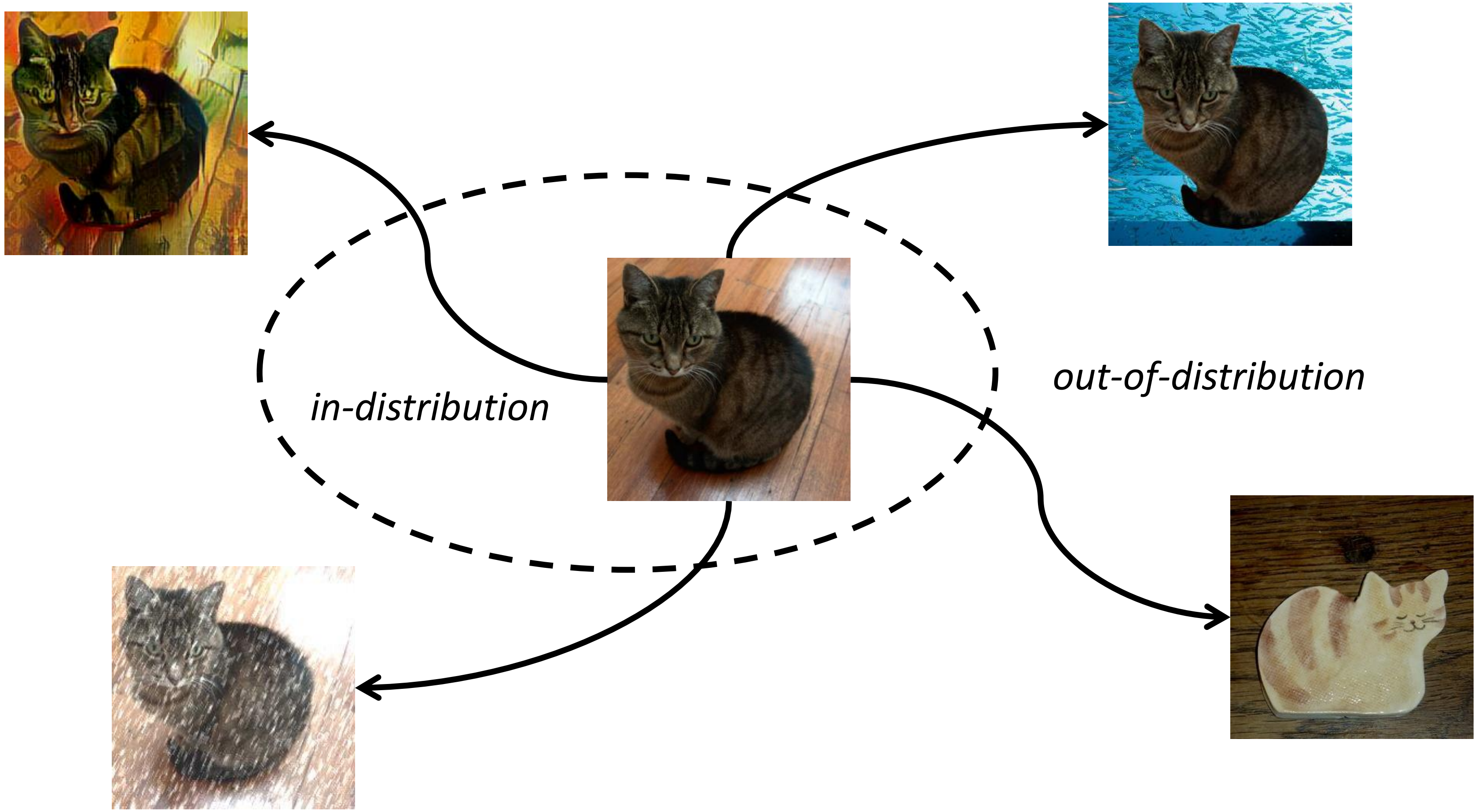
## Success of Vision Transformers





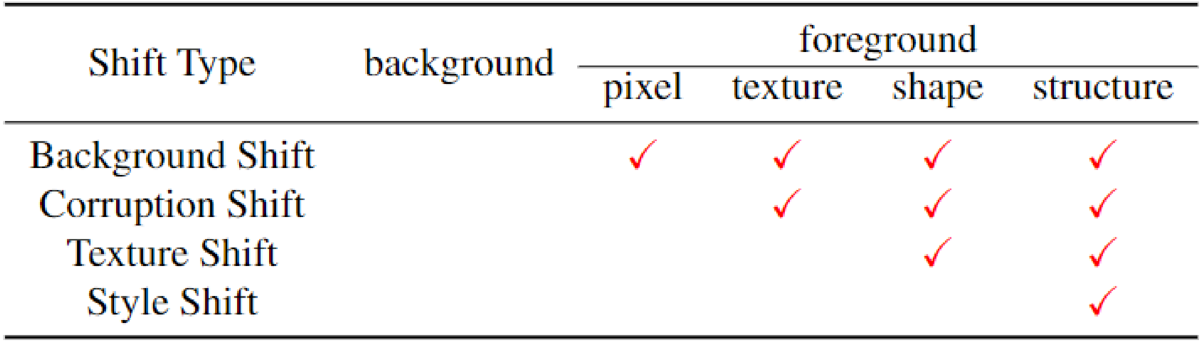
# 2D Sensory Data with Distribution Shifts

Taxonomy of out-of-distribution shifts in 2D images



# Investigation Protocol

## Categorization of distribution shifts



| Shift Type | background | foreground | | | |
|---|---|---|---|---|---|
| | | pixel | texture | shape | structure |
| Background Shift | | ✓ | ✓ | ✓ | ✓ |
| Corruption Shift | | | ✓ | ✓ | ✓ |
| Texture Shift | | | | ✓ | ✓ |
| Style Shift | | | | | ✓ |

## Out-of-distribution (OOD) generalization evaluation protocols

- ➢ Accuracy on OOD Data

$$Acc(F, C; \mathcal{D}_{ood}) = \frac{1}{|\mathcal{D}_{ood}|} \sum_{(\mathrm{x}, y) \in \mathcal{D}_{ood}} \mathbf{1}\,(C(F(\mathrm{x})) = y).$$

- ➢ IID/OOD Generalization Gap

$$Gap(F, C; \mathcal{D}_{iid}, \mathcal{D}_{ood}) = Acc(F, C; \mathcal{D}_{iid}) - Acc(F, C; \mathcal{D}_{ood}).$$

# Experimental Results and Analysis

## Background shift results

- *ViTs perform with a weaker background-bias than CNNs.*
- *A larger ViT extracts a more background-irrelevant representation.*

ImageNet-9 Dataset


Original

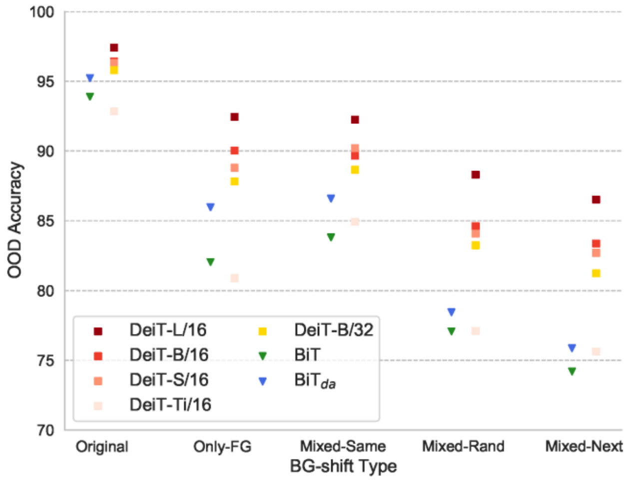

Only-FG


Mixed-Next

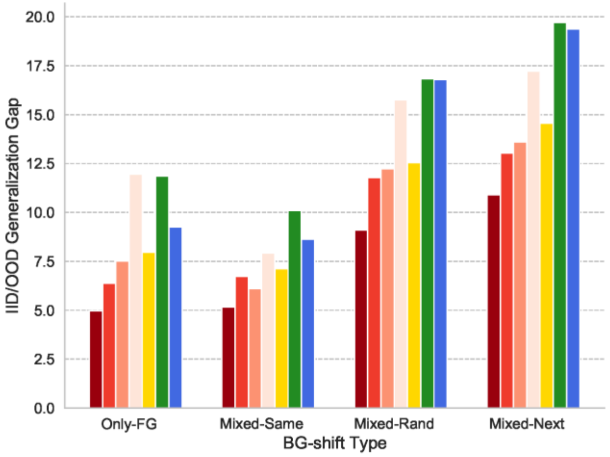

Mixed-Rand


Mixed-Same

ImageNet-9 Results



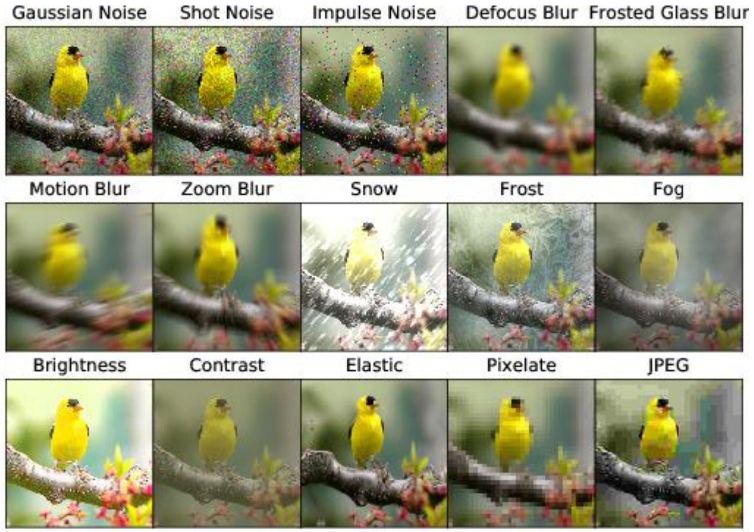
OOD Accuracy



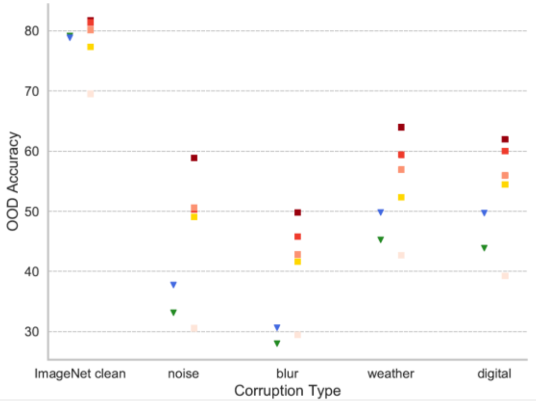
IID/OOD Generalization Gap

# Experimental Results and Analysis

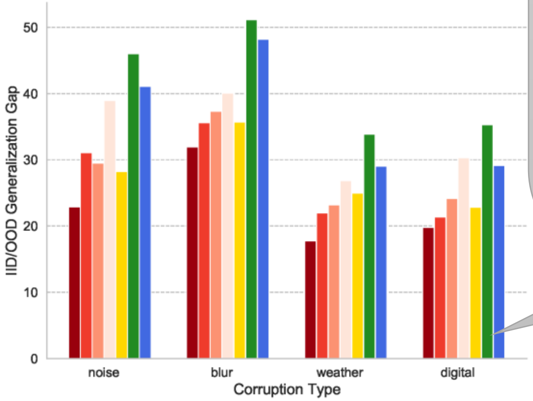
## Corruption shift results



ImageNet-C Dataset



OOD Accuracy



IID/OOD Generalization Gap

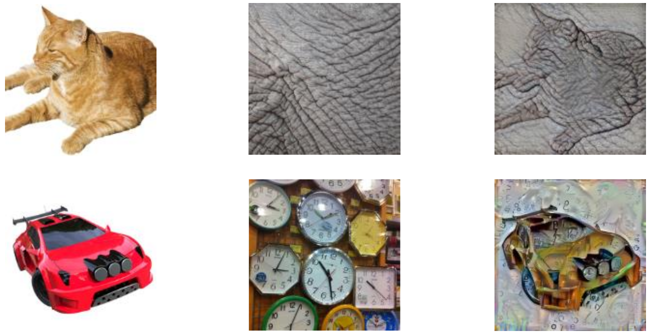
ImageNet-C Results

- *ViTs deal with corruption shifts better than CNNs and generalize better along with model size scaling up.*
- *ViTs benefit from diverse augmentation in enhancing generalization towards vicinal impurities, but their architectural advantage cannot be overlooked.*

# Experimental Results and Analysis

## Texture shift results

Stylized-ImageNet Dataset

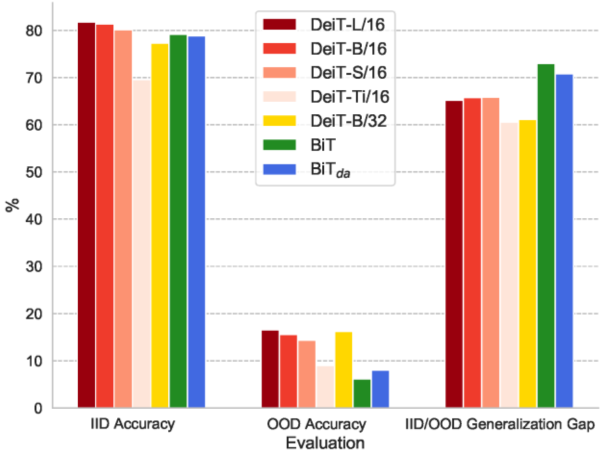


Context

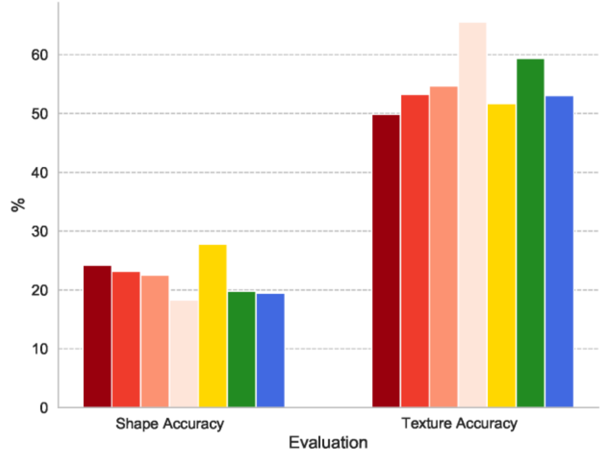
Texture

Cue Conflict Stimuli

Cue Conflict Stimuli Dataset



Stylized-ImageNet Results

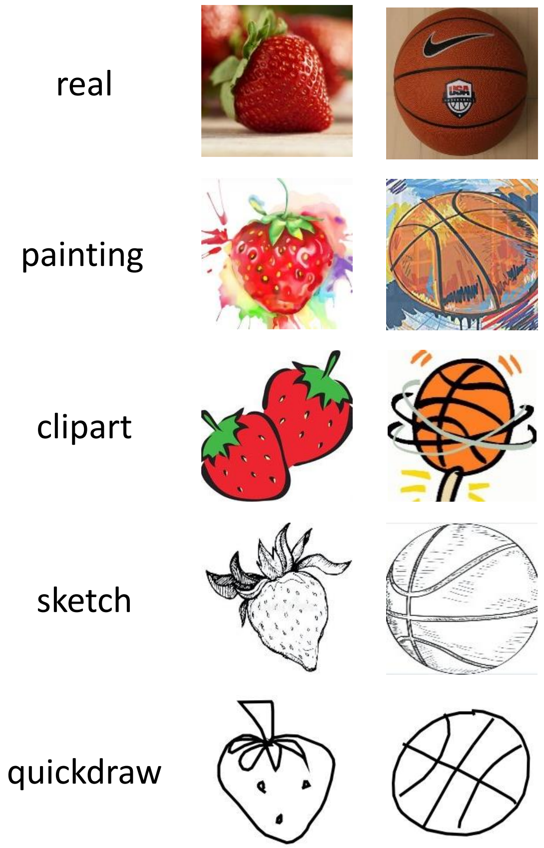


Cue Conflict Stimuli Results

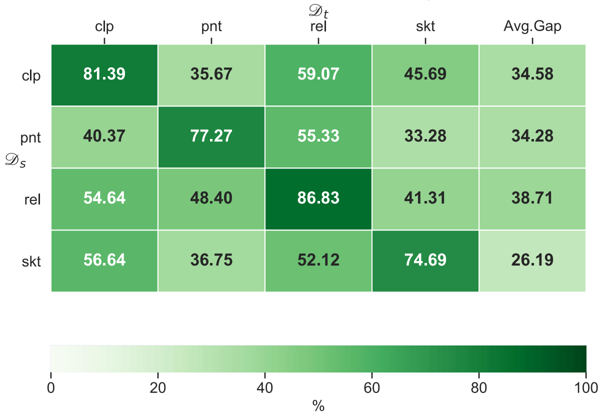
- *ViTs' stronger bias towards shape enables them to generalize better under texture shifts and their shape biases have a positive correlation with their sizes.*
- *ViTs with larger patch size exhibit a stronger bias towards the shape.*

# Experimental Results and Analysis

## Style shift results

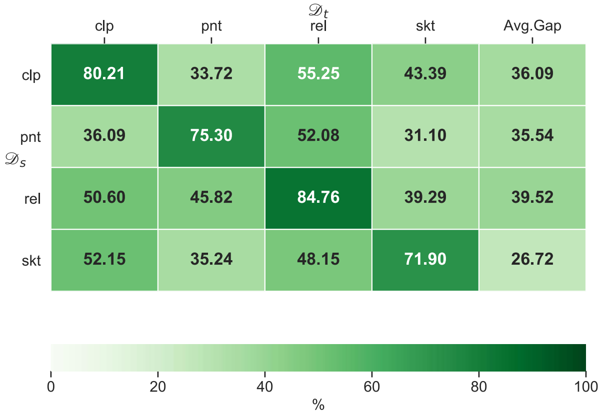


DomainNet Dataset



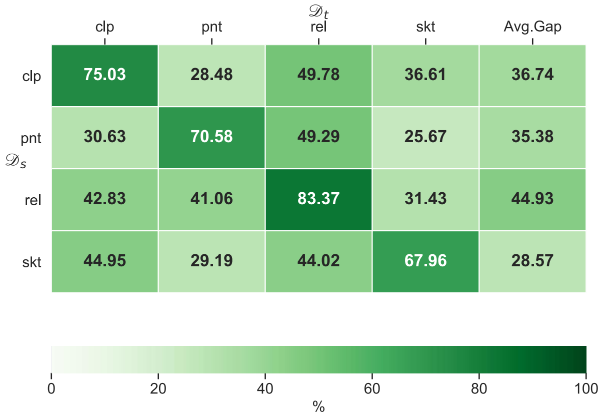
|  | clp | pnt | rel | skt | Avg.Gap |
|---|---|---|---|---|---|
| clp | 81.39 | 35.67 | 59.07 | 45.69 | 34.58 |
| pnt | 40.37 | 77.27 | 55.33 | 33.28 | 34.28 |
| rel | 54.64 | 48.40 | 86.83 | 41.31 | 38.71 |
| skt | 56.64 | 36.75 | 52.12 | 74.69 | 26.19 |

DeiT-B/16



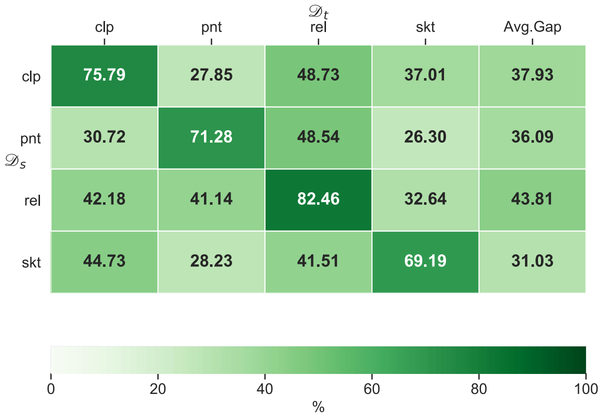
|  | clp | pnt | rel | skt | Avg.Gap |
|---|---|---|---|---|---|
| clp | 80.21 | 33.72 | 55.25 | 43.39 | 36.09 |
| pnt | 36.09 | 75.30 | 52.08 | 31.10 | 35.54 |
| rel | 50.60 | 45.82 | 84.76 | 39.29 | 39.52 |
| skt | 52.15 | 35.24 | 48.15 | 71.90 | 26.72 |

DeiT-S/16



|  | clp | pnt | rel | skt | Avg.Gap |
|---|---|---|---|---|---|
| clp | 75.03 | 28.48 | 49.78 | 36.61 | 36.74 |
| pnt | 30.63 | 70.58 | 49.29 | 25.67 | 35.38 |
| rel | 42.83 | 41.06 | 83.37 | 31.43 | 44.93 |
| skt | 44.95 | 29.19 | 44.02 | 67.96 | 28.57 |

BiT



|  | clp | pnt | rel | skt | Avg.Gap |
|---|---|---|---|---|---|
| clp | 75.79 | 27.85 | 48.73 | 37.01 | 37.93 |
| pnt | 30.72 | 71.28 | 48.54 | 26.30 | 36.09 |
| rel | 42.18 | 41.14 | 82.46 | 32.64 | 43.81 |
| skt | 44.73 | 28.23 | 41.51 | 69.19 | 31.03 |

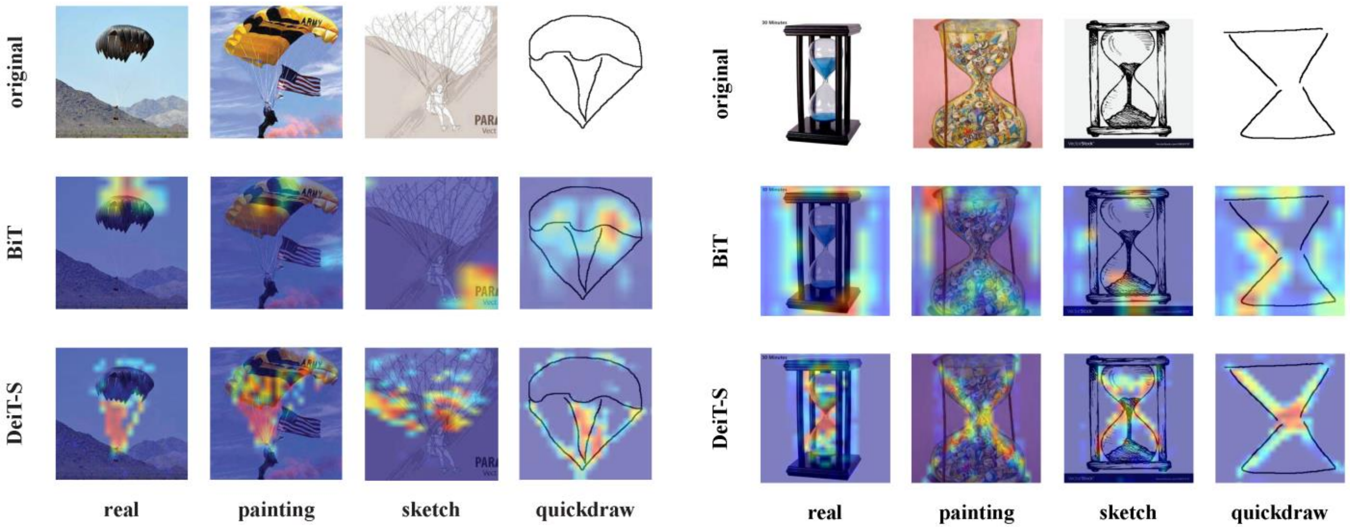
$\text{BiT}_{da}$

DomainNet Results

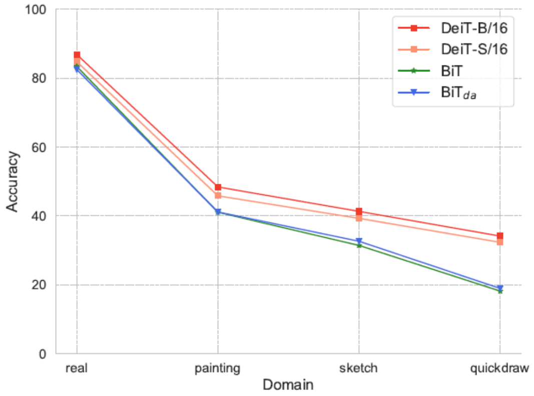
- *ViTs have diverse performance on IID/OOD generalization gap under Style shifts.*

# Experimental Results and Analysis

## Structure bias investigation



Grad-CAM Heat Maps

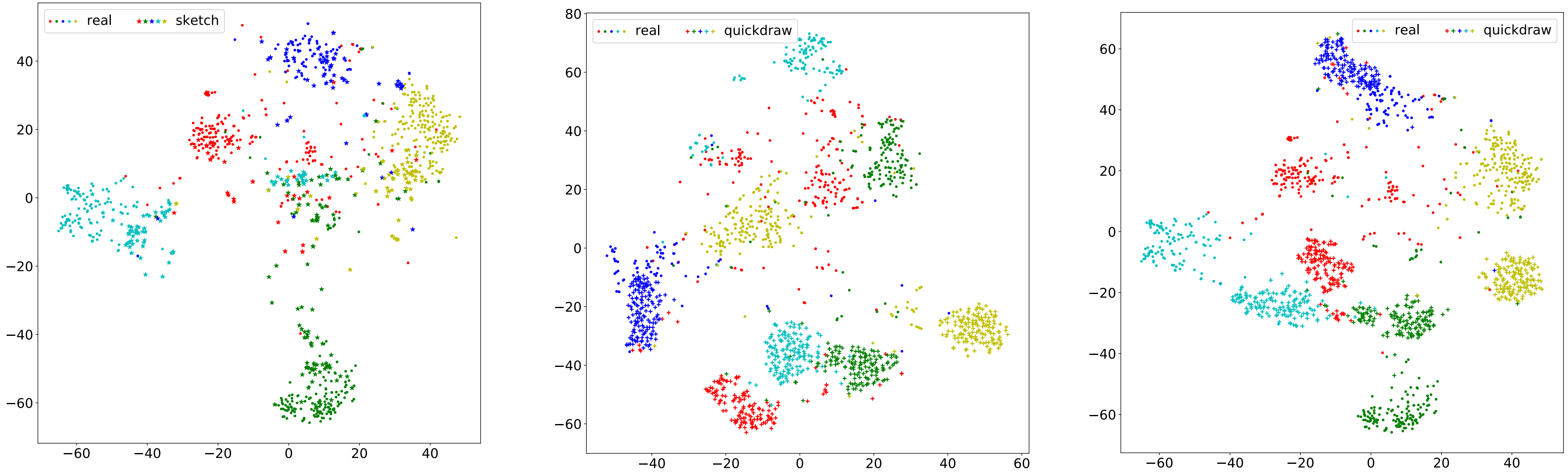


Accuracies of models trained with real on different domains

- *ViTs shows stronger bias towards object structure.*

# Experimental Results and Analysis

## Structure bias investigation



real vs. painting

real vs. sketch

real vs. quickdraw

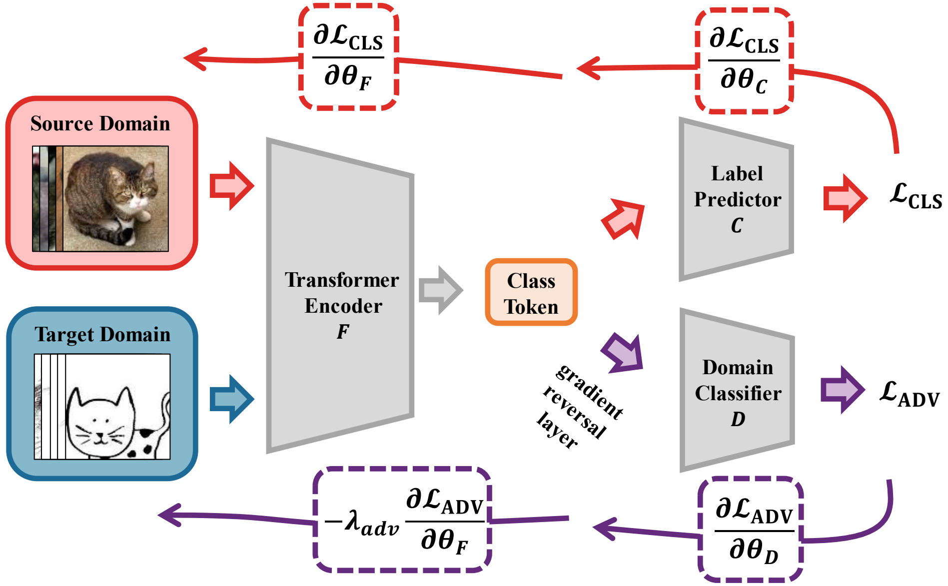
- *ViTs will eliminate different levels of DS in different layers.*

T-SNE Visualization Results in Layer 12

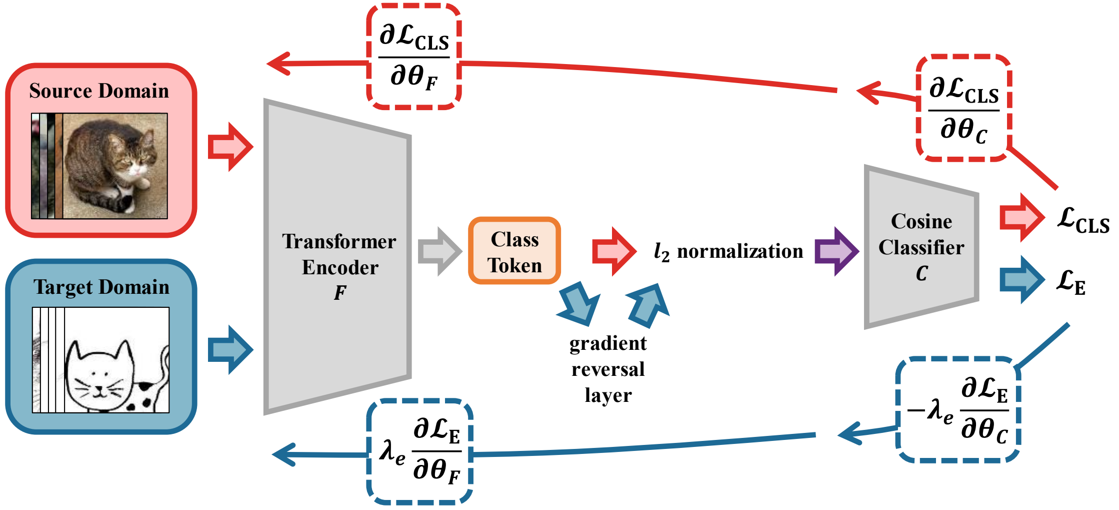
# Enhancing Generalization of ViTs

## Generalization-Enhanced ViTs

### T-ADV (based on adversarial learning)



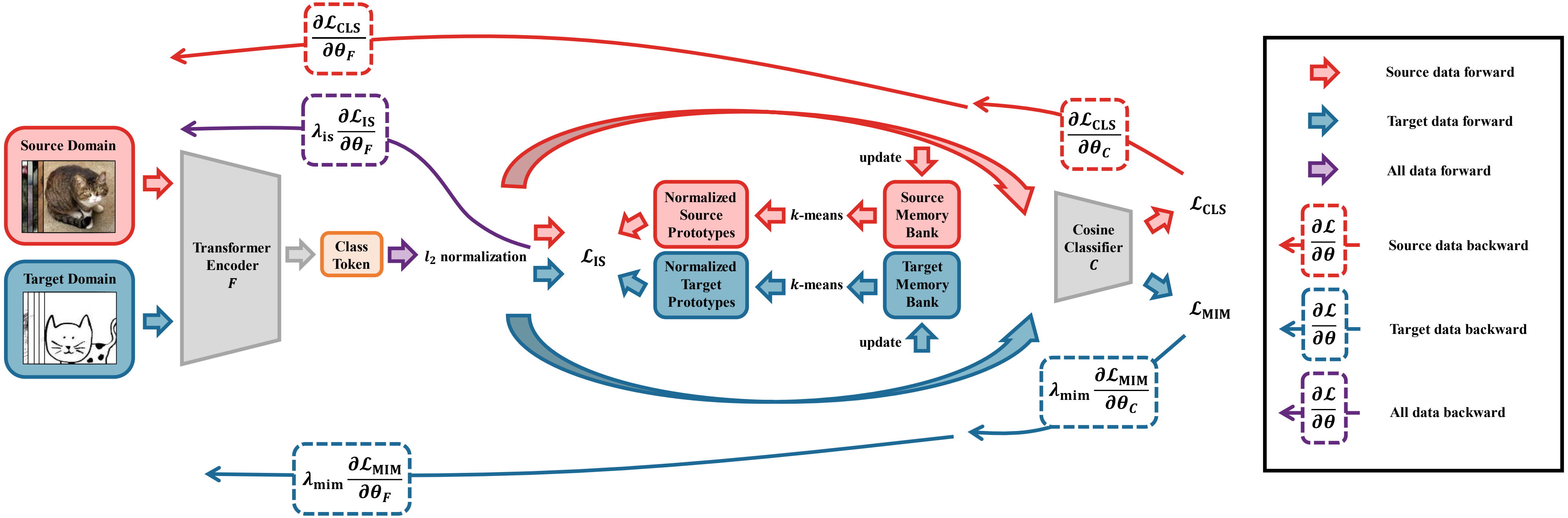
### T-MME (based on minimax entropy)



# Enhancing Generalization of ViTs

## Generalization-Enhanced ViTs

## T-SSL (based on self-supervised learning)

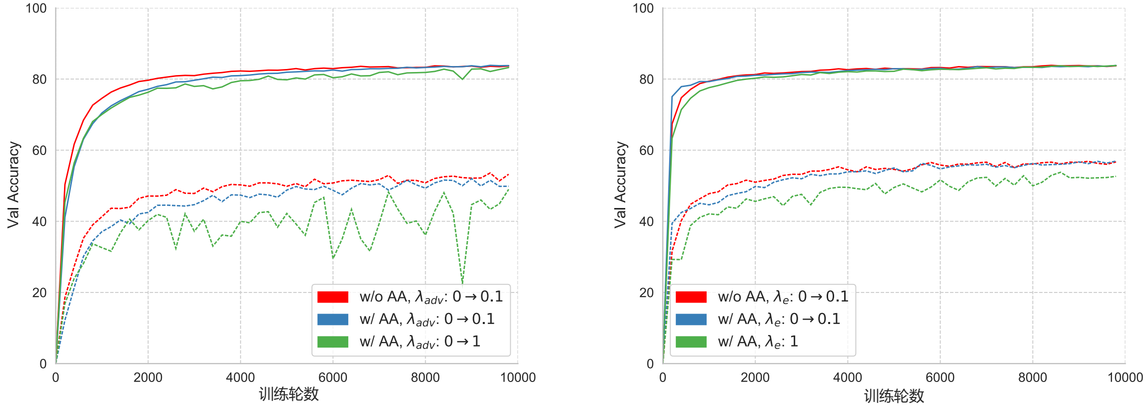


# Enhancing Generalization of ViTs

## Studies on Generalization-Enhanced ViTs

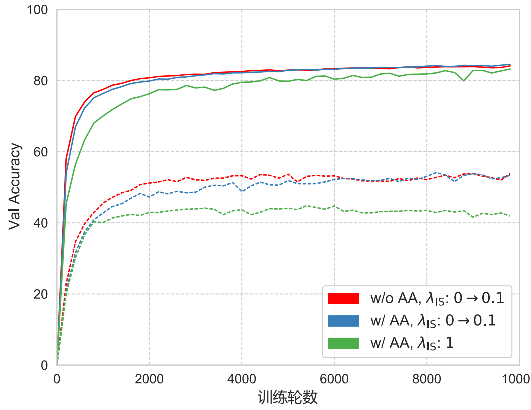
| Model | Method | R→C | R→P | P→C | C→S | S→P | R→S | P→R | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| **DeiT-B/16** | - | 54.6 | 48.4 | 40.4 | 45.7 | 36.8 | 41.3 | 55.3 | 46.1 |
| | T-ADV | 58.2 | 50.9 | 41.9 | 51.2 | 46.1 | 47.5 | 55.7 | 50.2 |
| | T-MME | 60.6 | 52.0 | 42.3 | 50.3 | 45.8 | 48.0 | 54.9 | 50.5 |
| | T-SSL | 56.8 | 49.1 | 46.0 | 51.8 | 47.0 | 46.0 | 61.0 | 51.1 |
| **DeiT-S/16** | - | 50.6 | 45.8 | 36.1 | 43.4 | 35.2 | 39.3 | 52.1 | 43.2 |
| | T-ADV | 53.6 | 47.8 | 38.0 | 47.1 | 41.6 | 41.9 | 52.8 | 46.1 |
| | T-MME | 56.9 | 49.2 | 39.0 | 46.5 | 43.0 | 42.1 | 52.5 | 47.0 |
| | T-SSL | 53.9 | 46.7 | 42.8 | 47.3 | 43.0 | 40.9 | 57.1 | 47.4 |
| **BiT** | - | 42.2 | 41.1 | 30.7 | 37.0 | 28.2 | 32.6 | 48.5 | 36.8 |
| | DANN | 45.2 | 42.9 | 33.0 | 40.4 | 36.6 | 35.3 | 49.3 | 40.4 |
| | MME | 50.2 | 44.6 | 34.8 | 40.3 | 38.4 | 37.8 | 47.6 | 42.0 |
| | SSL | 52.6 | 42.8 | 39.0 | 45.7 | 39.1 | 39.7 | 56.1 | 45.0 |
| **VGG-16** | - | 39.4 | 37.3 | 26.4 | 33.0 | 25.6 | 27.8 | 45.7 | 33.6 |
| | DANN | 43.3 | 40.1 | 28.7 | 36.2 | 31.6 | 35.5 | 44.7 | 37.2 |
| | MME | 42.7 | 42.5 | 27.4 | 36.9 | 33.9 | 32.6 | 45.9 | 37.4 |
| | SSL | 43.8 | 41.9 | 32.2 | 35.7 | 37.0 | 31.1 | 55.2 | 39.5 |

Results of Generalization-enhanced methods



T-ADV

T-MME

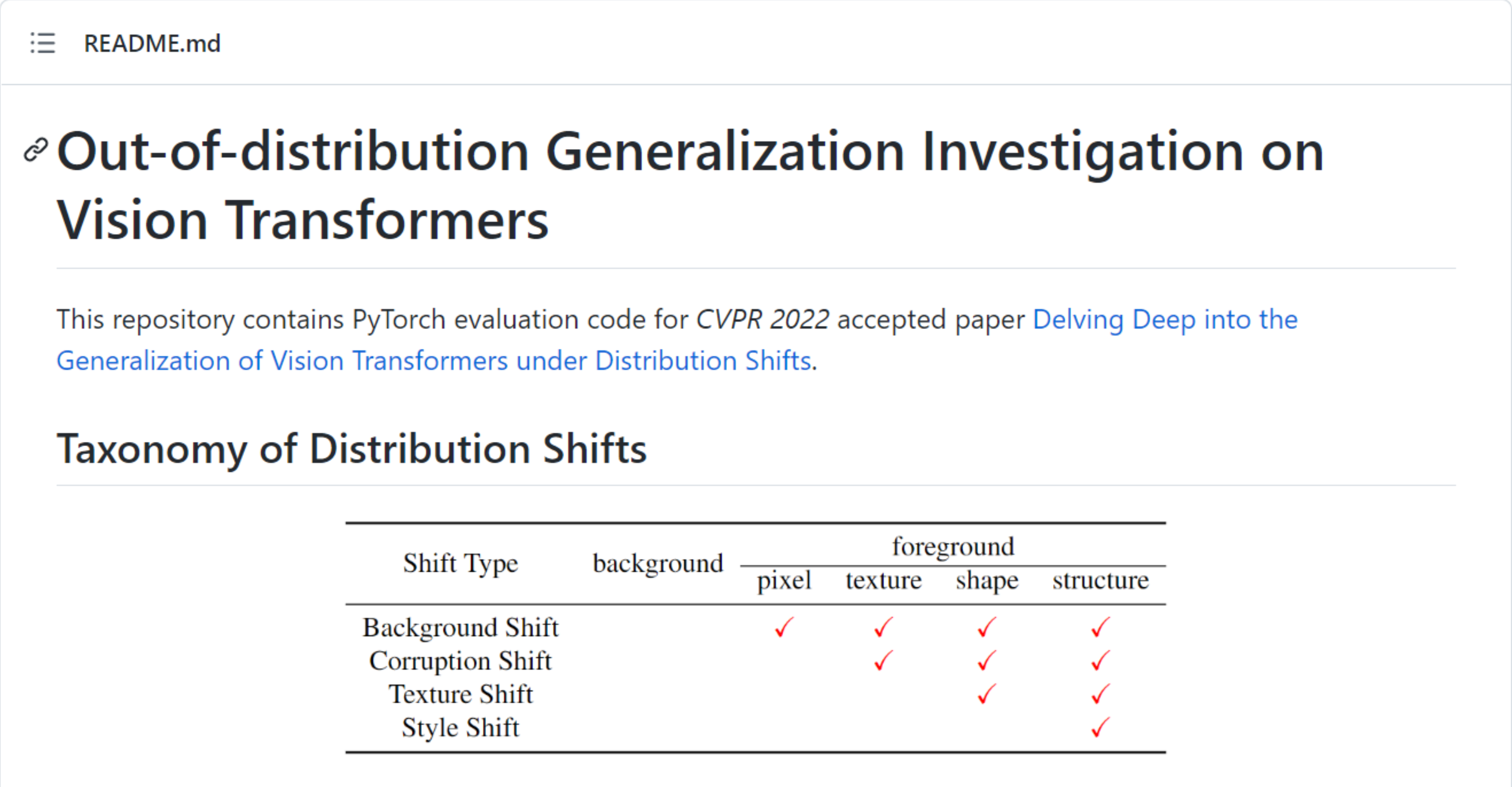


T-SSL

Effectiveness of different training strategies

# Code and models

- Released at [https://github.com/Phoenix1153/ViT_OOD_generalization](https://github.com/Phoenix1153/ViT_OOD_generalization)



README.md

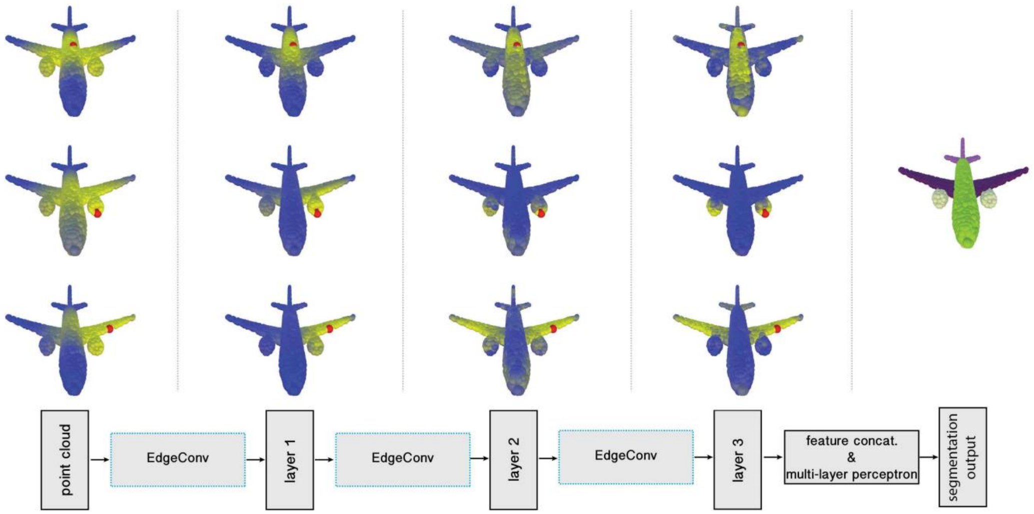
# Out-of-distribution Generalization Investigation on Vision Transformers

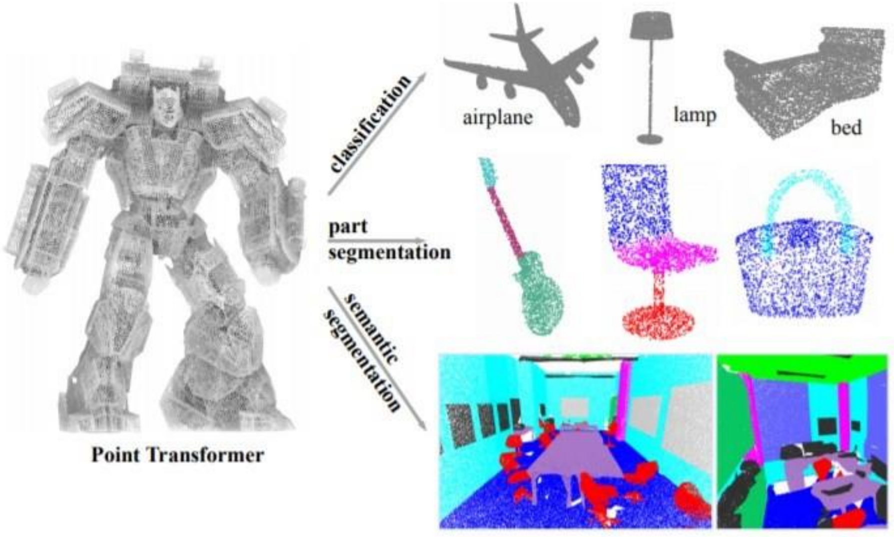
This repository contains PyTorch evaluation code for *CVPR 2022* accepted paper Delving Deep into the Generalization of Vision Transformers under Distribution Shifts.

## Taxonomy of Distribution Shifts

| Shift Type | background | foreground | | | |
|---|---|---|---|---|---|
| | | pixel | texture | shape | structure |
| Background Shift | | ✓ | ✓ | ✓ | ✓ |
| Corruption Shift | | | ✓ | ✓ | ✓ |
| Texture Shift | | | | ✓ | ✓ |
| Style Shift | | | | | ✓ |

# Convolution v.s. Attention (3D Vision)





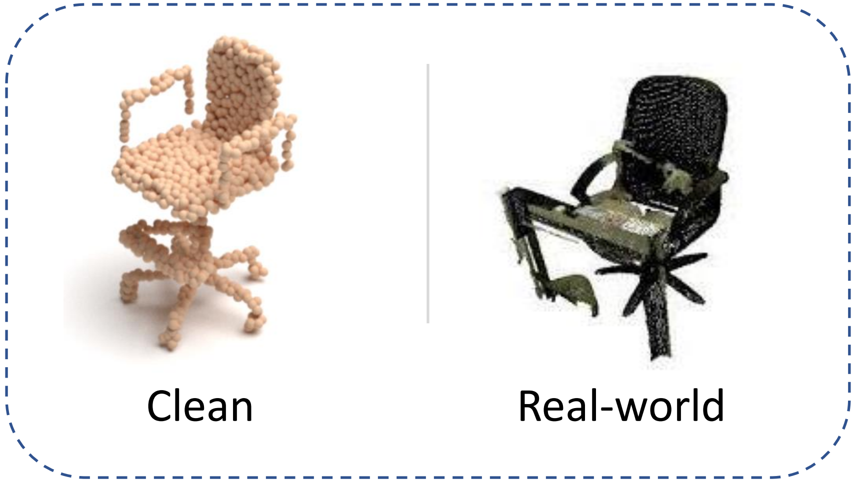
Ren et al., Benchmarking and Analyzing Point Cloud Classification under Corruptions, ArXiv 2022

Related Works:

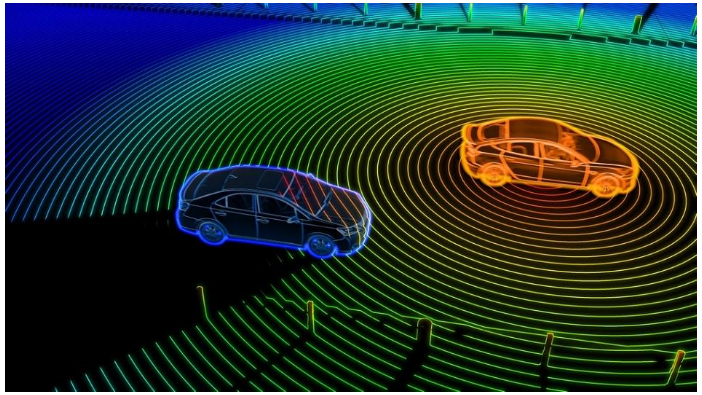
Sun et al., Benchmarking Robustness of 3D Point Cloud Recognition Against Common Corruptions, ArXiv 2022

# Robustness is Crucial in Point Cloud

- Point clouds are used in **safety-critical** applications but often suffer from severe **OOD corruptions**.



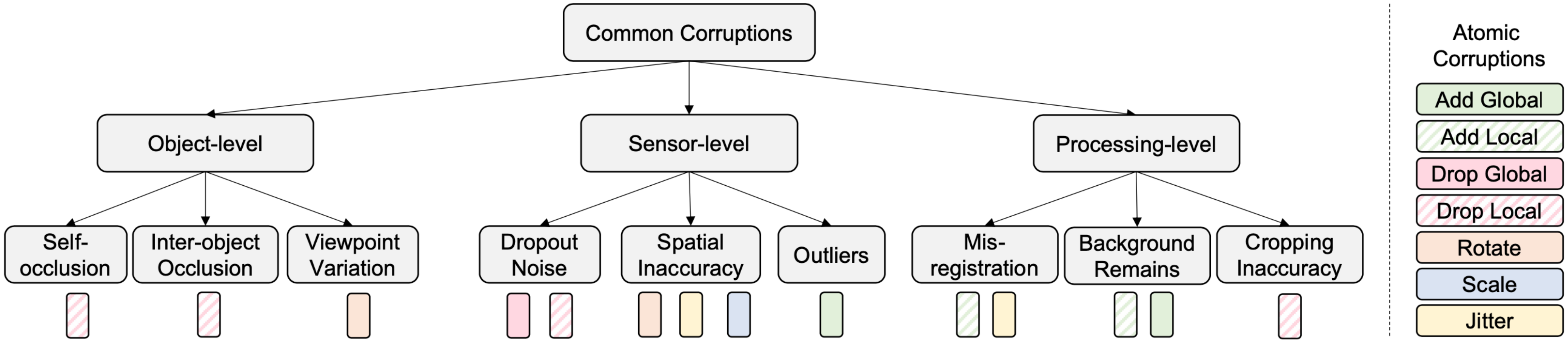
Corruptions are severe and OOD
e.g., occlusions, sensory noise



Applications are safety-critical
e.g., autonomous driving

# 3D Sensory Data with Distribution Shifts

- **Corruptions Taxonomy**: We break down common corruptions into detailed corruption sources, and further simplify them into a combination of atomic corruptions.



Common Corruptions

Object-level

- Self-occlusion
- Inter-object Occlusion
- Viewpoint Variation

Sensor-level

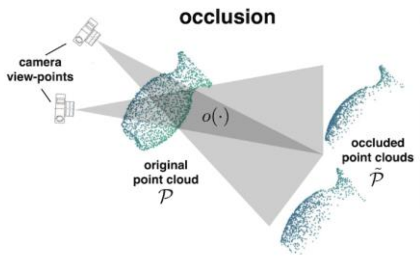
- Dropout Noise
- Spatial Inaccuracy
- Outliers

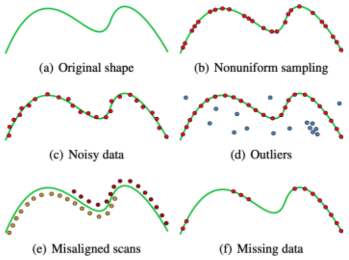
Processing-level

- Mis-registration
- Background Remains
- Cropping Inaccuracy

Atomic Corruptions

- Add Global
- Add Local
- Drop Global
- Drop Local
- Rotate
- Scale
- Jitter

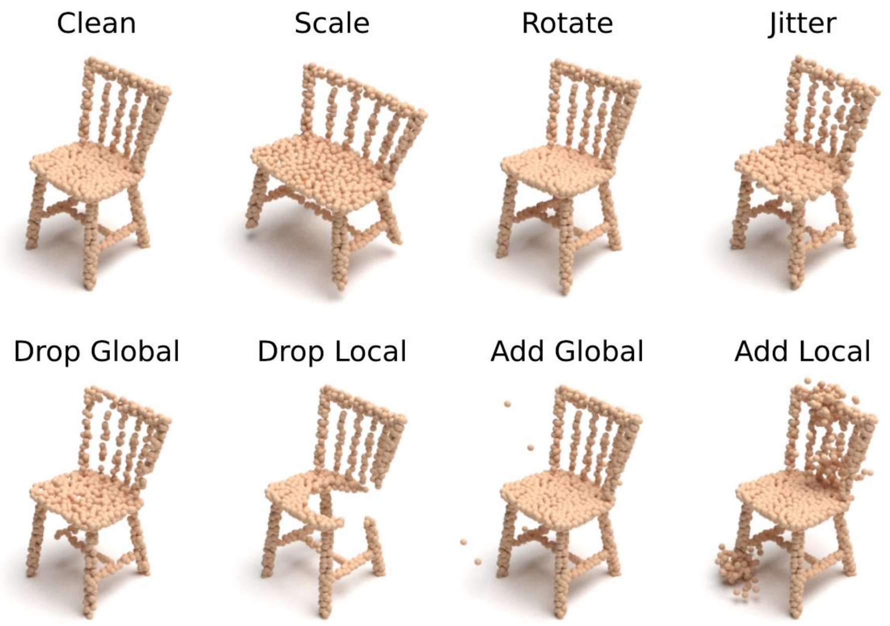




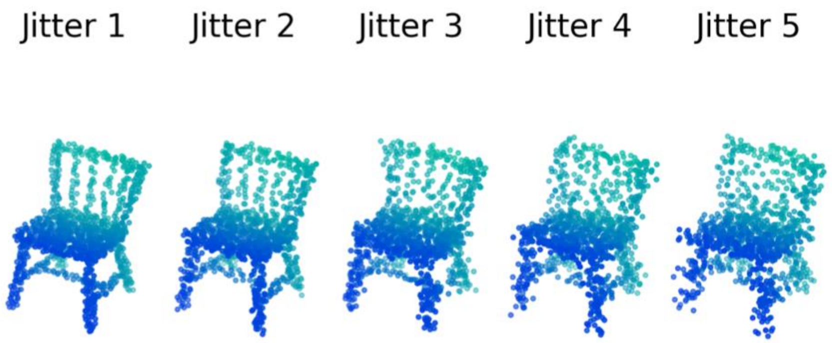
(a) OBJ_BG (b) PB_T25 (c) PB_T25_R

# Comprehensive Benchmarking Suite

**ModelNet-C**: ModelNet40 is one of the most used benchmarks. We corrupt the ModelNet40 testset using the atomic corruptions with varying severities.



Atomic Corruptions



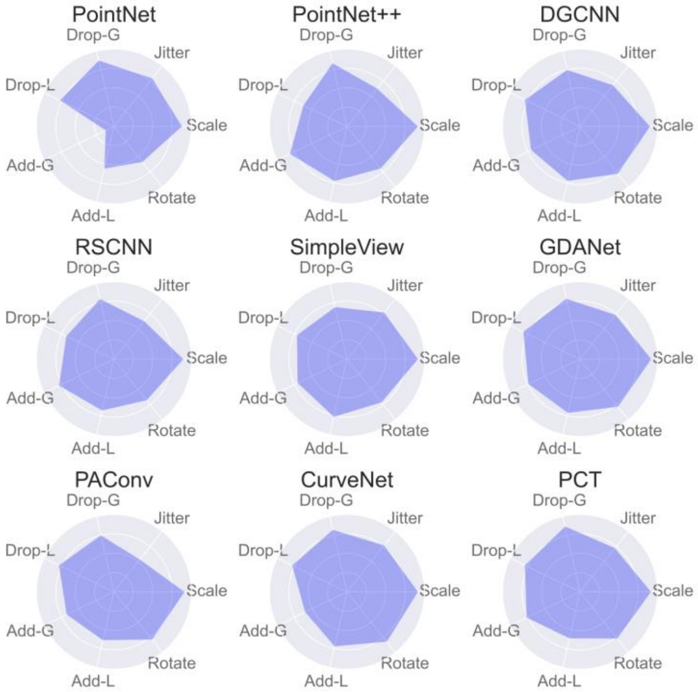
Different Severities

# Evaluation Protocol

**Evaluation Metrics**: Inspired by the ImageNet-C, we use mean CE (mCE), as the primary metric. Compared to the commonly used Overall Accuracy (OA), mCE shows average performance under all types of corruptions.

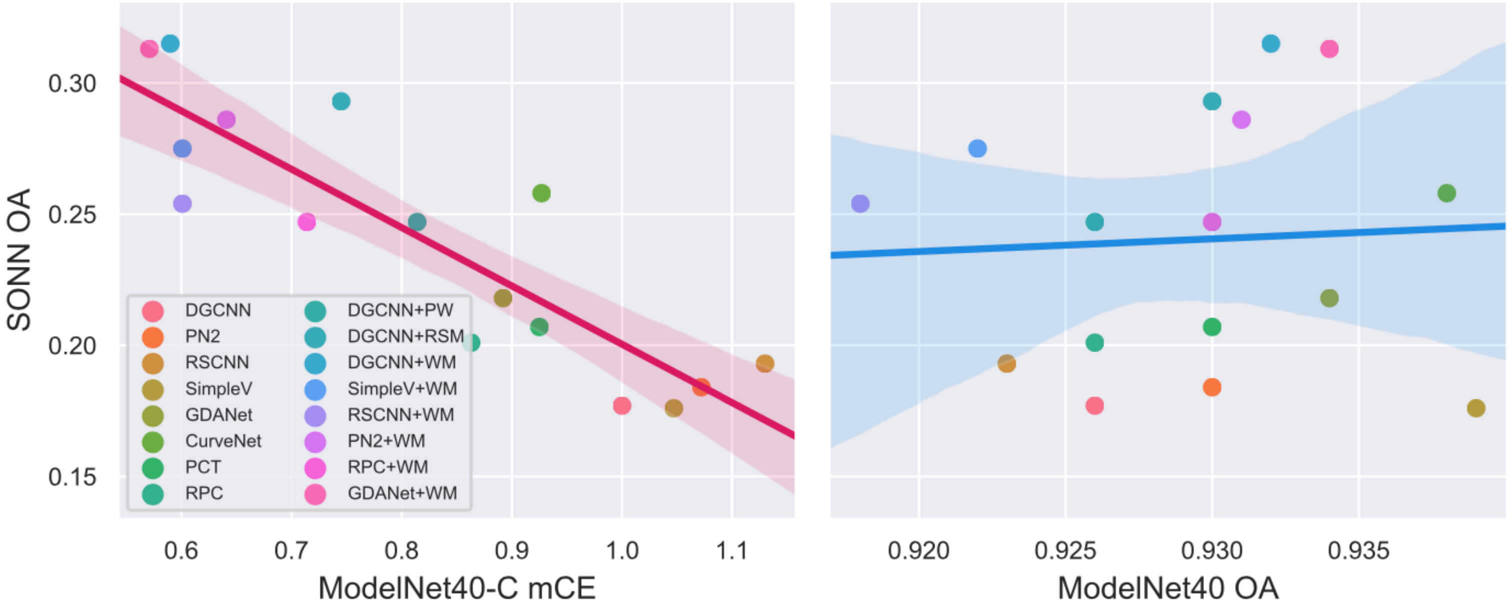
$$\mathrm{CE}_i = \frac{\sum_{l=1}^{5}(1 - \mathrm{OA}_{i,l})}{\sum_{l=1}^{5}(1 - \mathrm{OA}_{i,l}^{\mathrm{DGCNN}})},$$

$$\mathrm{mCE} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{CE}_i$$



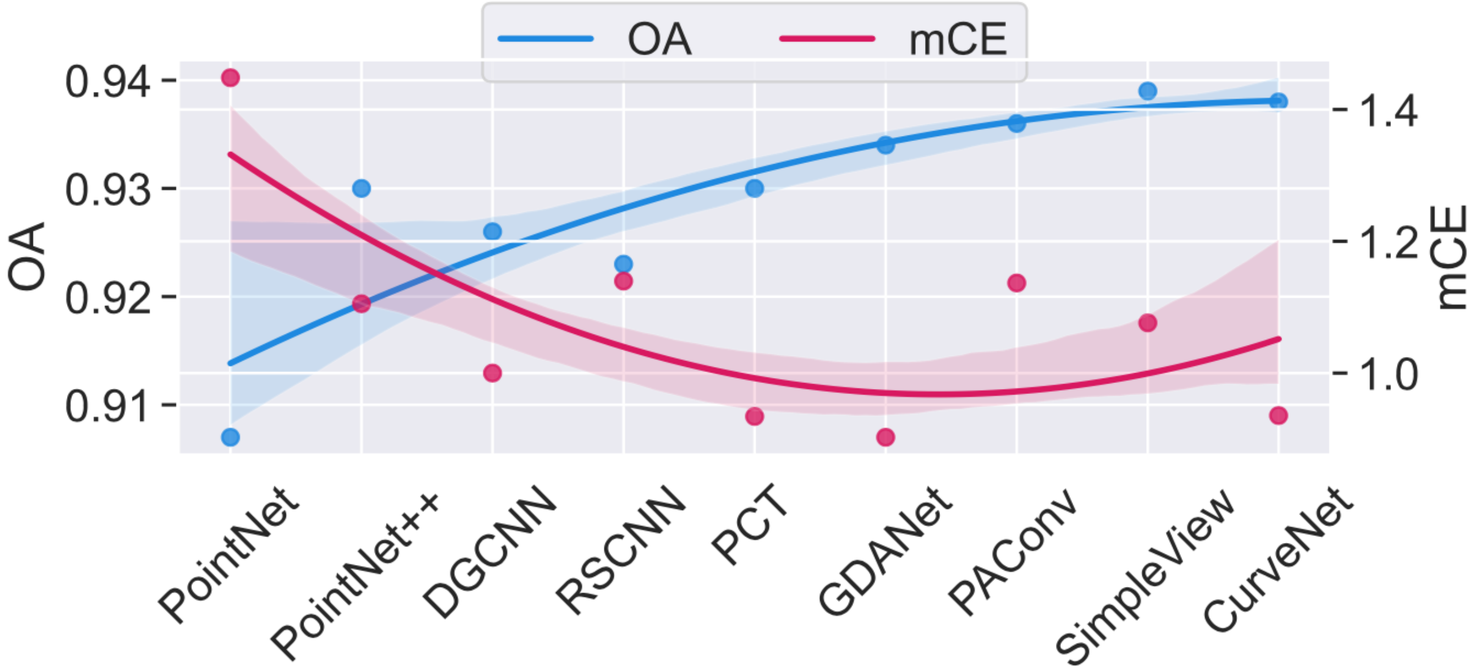
# Indicative of real-world robustness?

- **Yes.** We observe that ModelNet-C mCE strongly correlates to ScanObjectNN (SONN) OA. In comparison, ModelNet40 OA has nearly no correlation to SONN OA.



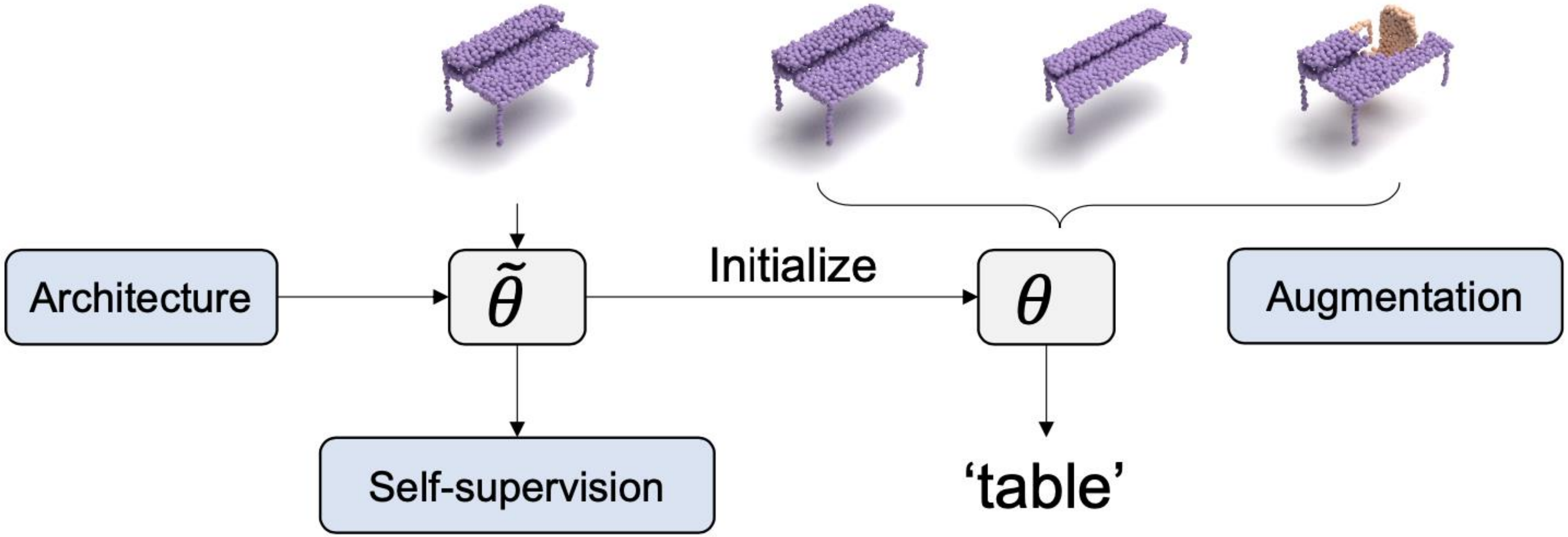
# Point cloud classifier getting more robust?

- **No.** Although the accuracy on ModelNet40 gradually saturates, the robustness is at the risk of getting worse, due to the lack of a standard test suite.



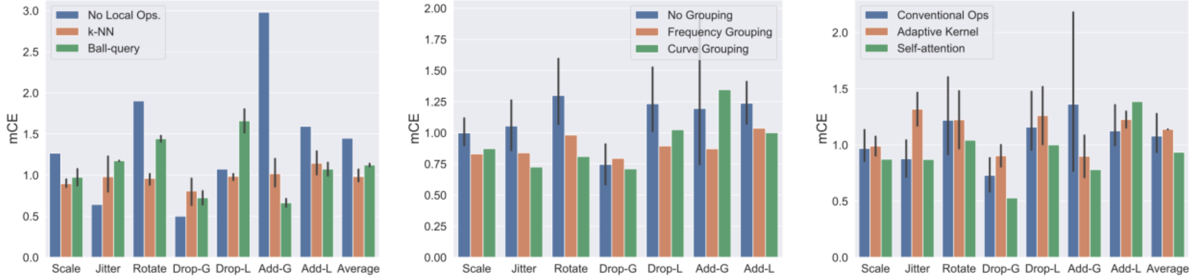
# What makes a robust point cloud classifier?

- **Three main components**: 1) architecture design, 2) self- supervised pretraining 3) augmentation methods.

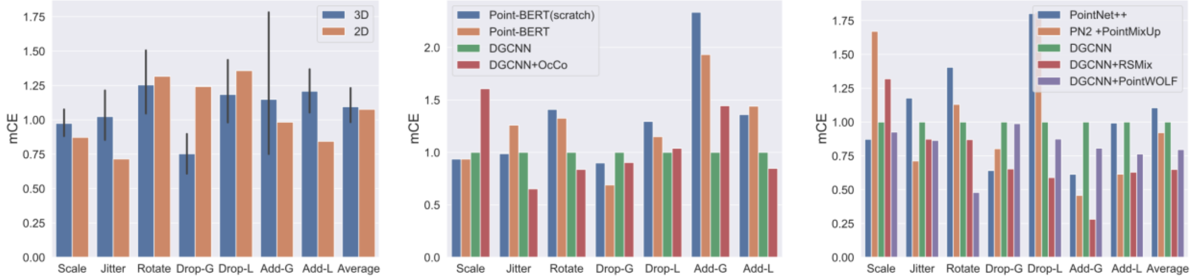


# What makes a robust point cloud classifier?

- We conduct a comprehensive analysis and observe:
  - Proper architecture designs can improve robustness, e.g., advanced grouping and self-attention.
  - Pretrain signals can be transferred, benefiting robustness under specific corruptions.
  - Mixing and deformation augmentations can bring significant improvements to model robustness.



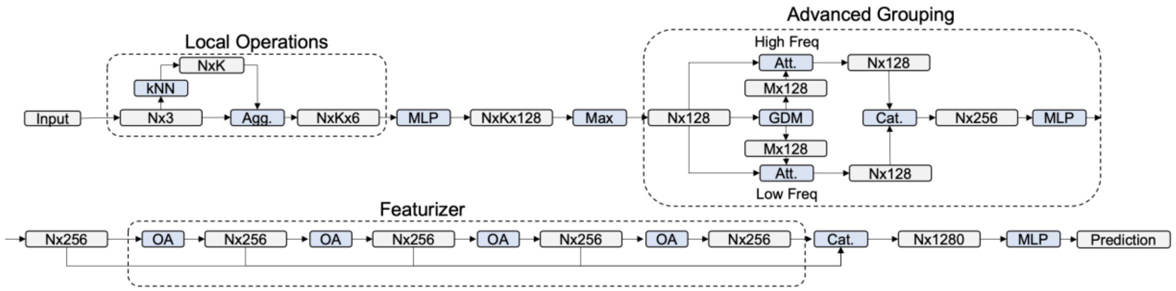
(a) Local Operations (b) Advanced Grouping (c) Featurizer



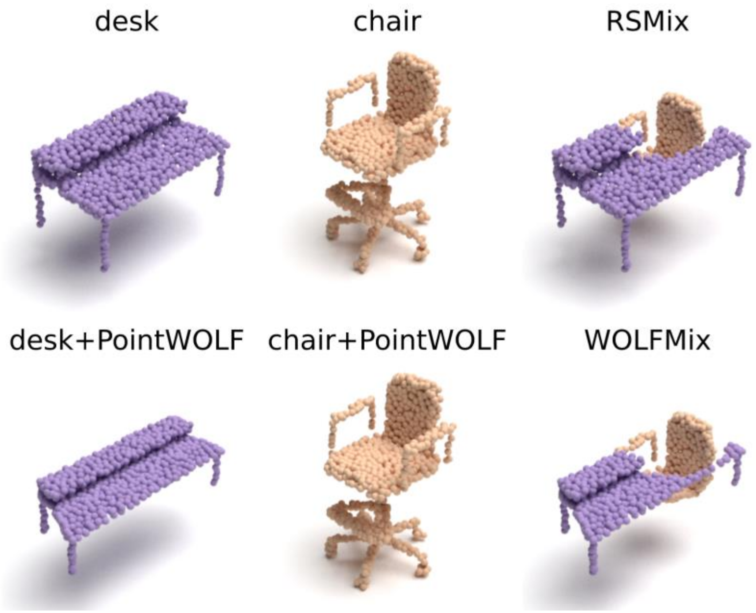
(d) 2D v.s. 3D (e) Pretrain (f) Augmentation

# Enhancing Robustness in Point Cloud

- For verification, we propose a new architecture and a new augmentation technique strictly following our empirical findings.
- They *outperform* existing methods.



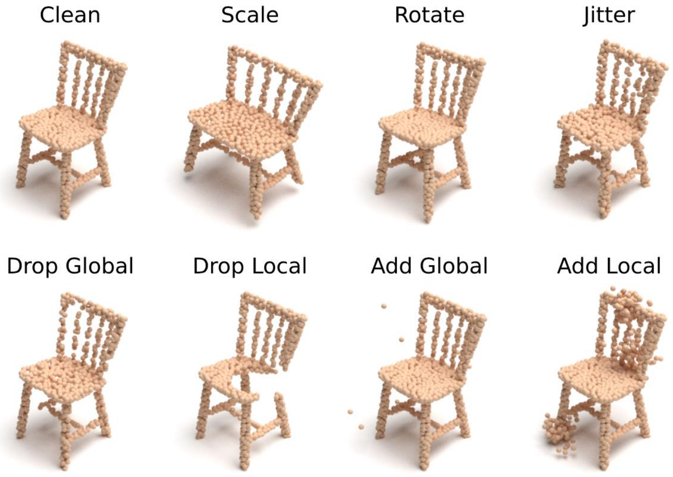
Our proposed architecture *RPC*



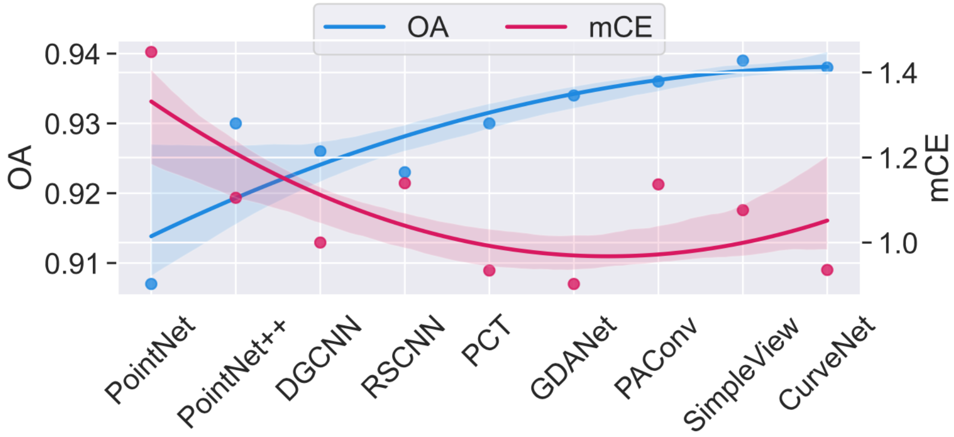
Our proposed augmentation *WolfMix*

# Conclusion

- The SoTA methods for point cloud classification on clean data are becoming **less robust** to random real-world corruptions.
- We highly encourage future research to **focus on classification robustness** so as to benefit real applications.



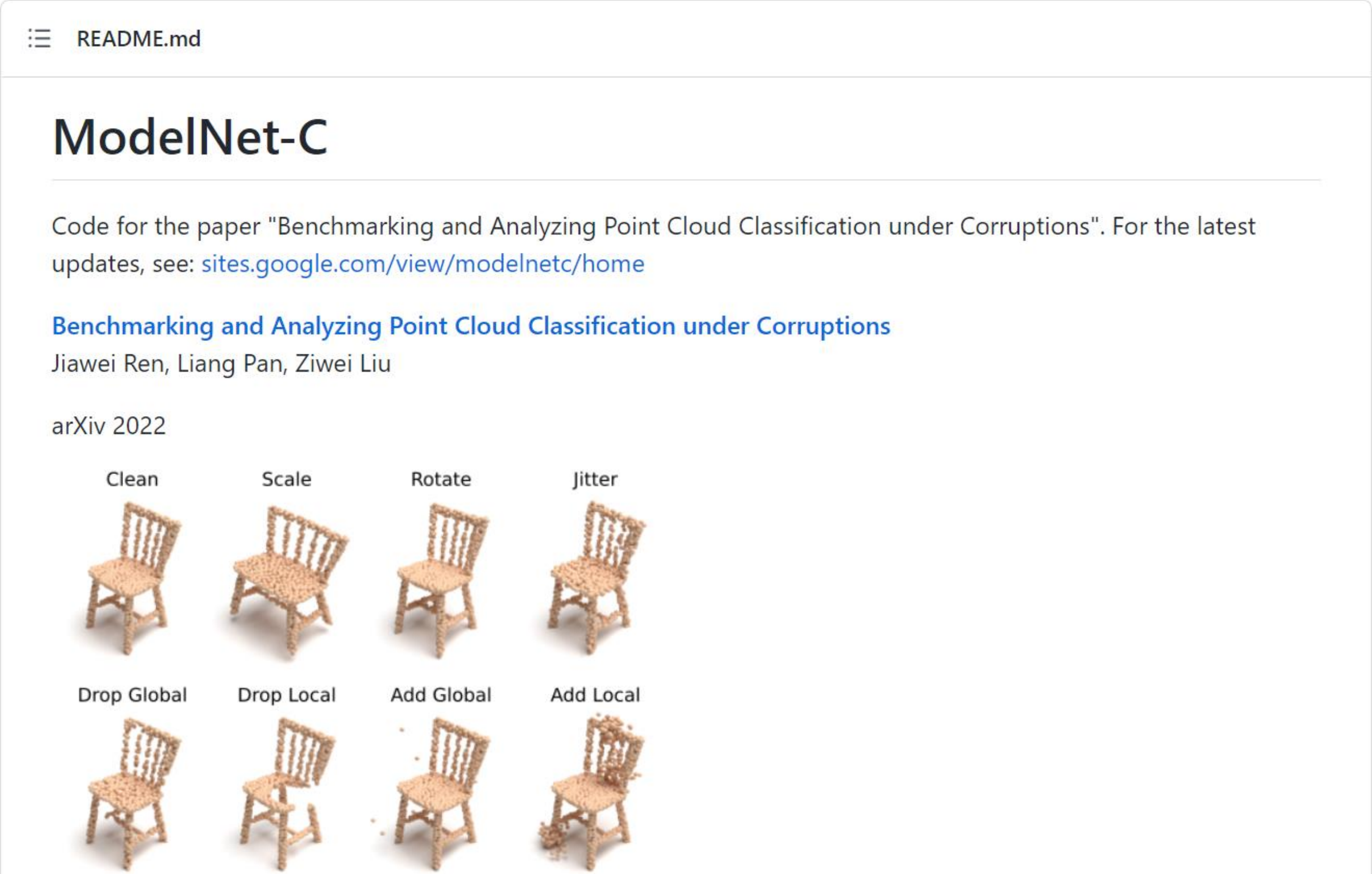
ModelNet-C



Benchmark Result

# Code, Models & Dataset

Released at https://github.com/jiawei-ren/ModelNet-C



## ModelNet-C

Code for the paper "Benchmarking and Analyzing Point Cloud Classification under Corruptions". For the latest updates, see: sites.google.com/view/modelnetc/home

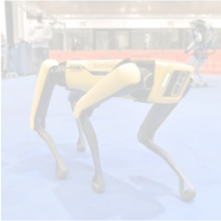
**Benchmarking and Analyzing Point Cloud Classification under Corruptions**
Jiawei Ren, Liang Pan, Ziwei Liu

arXiv 2022

# Generalization in Vision Models

**Semantic Shift**

*OOD Detection*



*Zero-shot / Few-shot / Long-tailed Learning*

**Sensory Modalities**

**Neural Architectures**

*Corruptions / Perturbations / Domain Shifts*

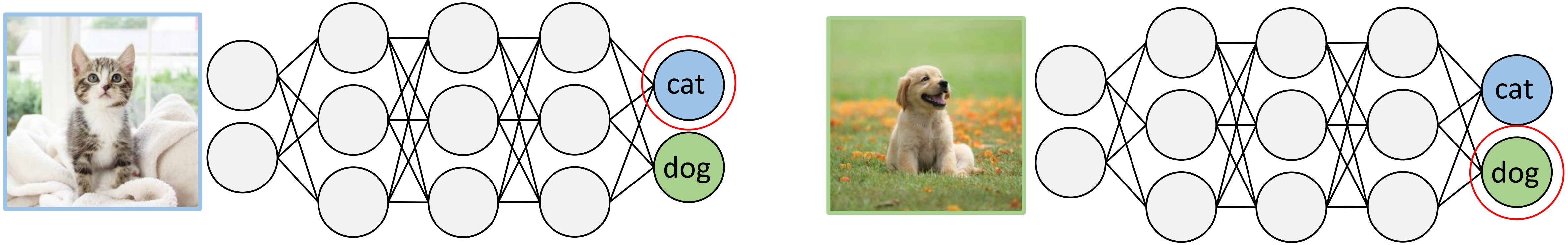
**Covariate Shift**

# Vision + Language

Zhou et al., Learning to Prompt for Vision-Language Models, ArXiv 2021
Zhou et al., Conditional Prompt Learning for Vision-Language Models, CVPR 2022

# Learning with discrete labels

- For image recognition we basically learn associations between images and discrete labels (represented by *randomly initialized vectors*)



# Problems with discrete labels

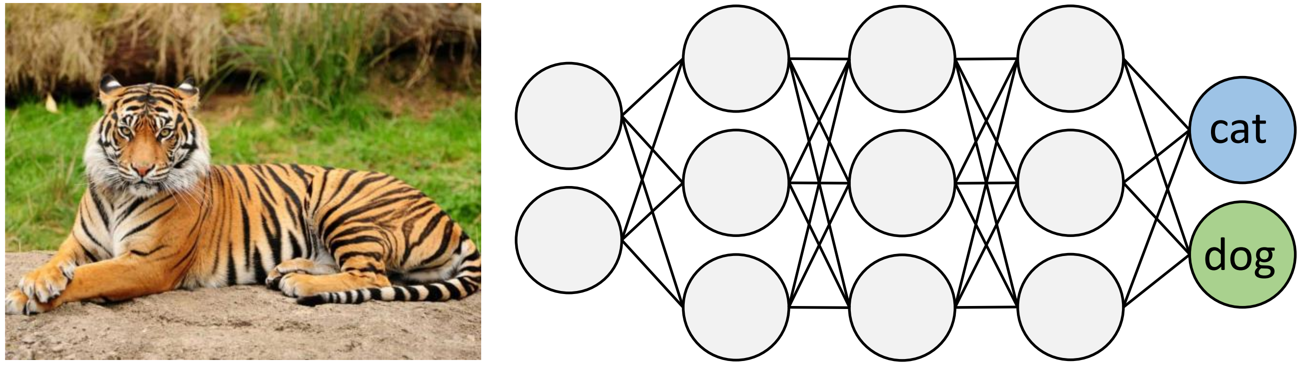
- Difficult to scale the dataset

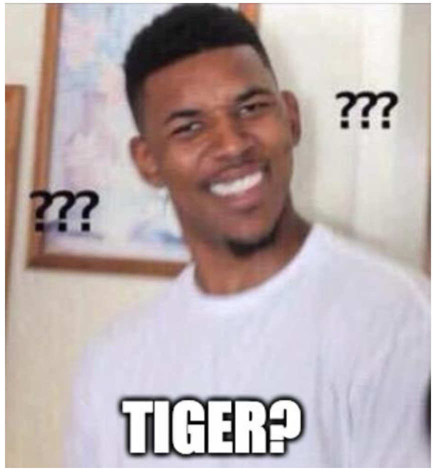
*We're talking about millions of images*

*Ambiguity: a baby or a cat?*

# Problems with discrete labels

- Cannot generalize to new concepts (new data needs to be collected)





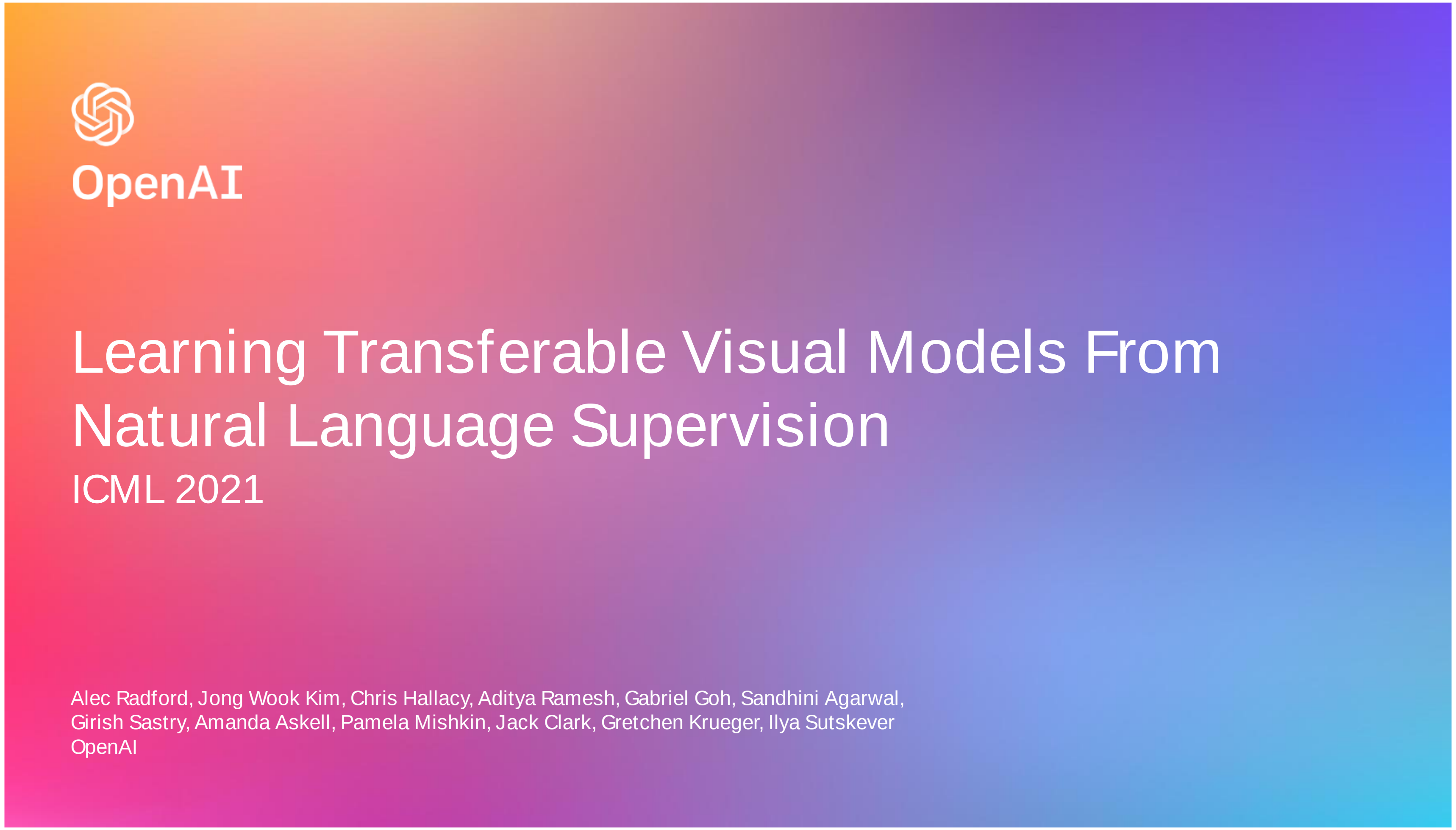
# Learning with multi-modality signals

- Using natural language as supervision

*Caption: a baby holding a kitten*

- more accurate description
- can easily scale up the dataset (just search image-text pairs or use image & alt-text)

# Large vision-language models



OpenAI

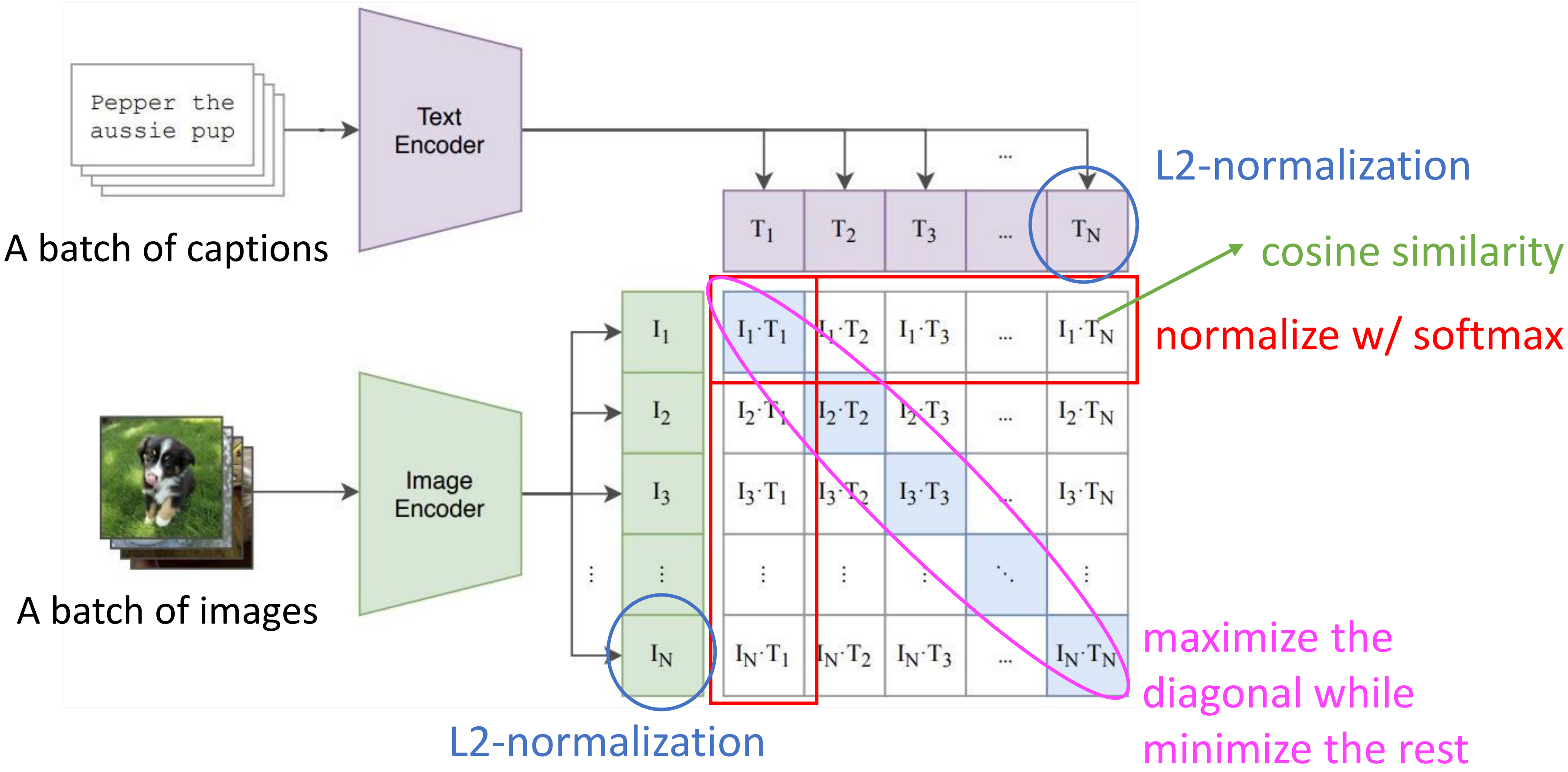
## Learning Transferable Visual Models From Natural Language Supervision

ICML 2021

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever
OpenAI

# Contrastive language-image pre-training

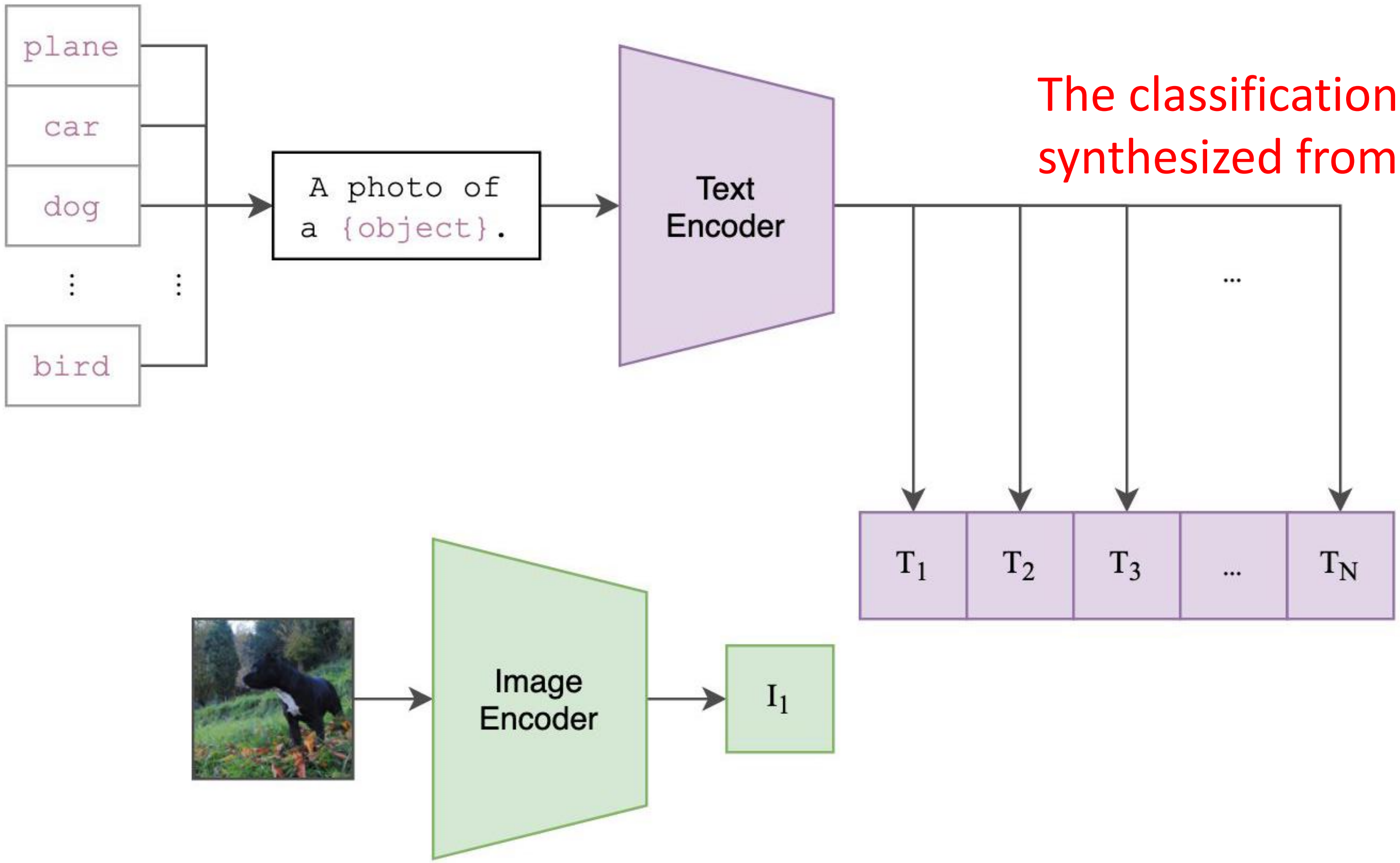
- Training pipeline



Radford et al. Learning transferable visual models from natural language supervision. ICML 2021.

# Contrastive language-image pre-training

- Test time: can naturally do zero-shot recognition



The classification weights (representing visual concepts) are synthesized from natural language, a.k.a. *prompting*

Radford et al. Learning transferable visual models from natural language supervision. ICML 2021.

# Remarkable zero-shot performance & robustness to domain shift

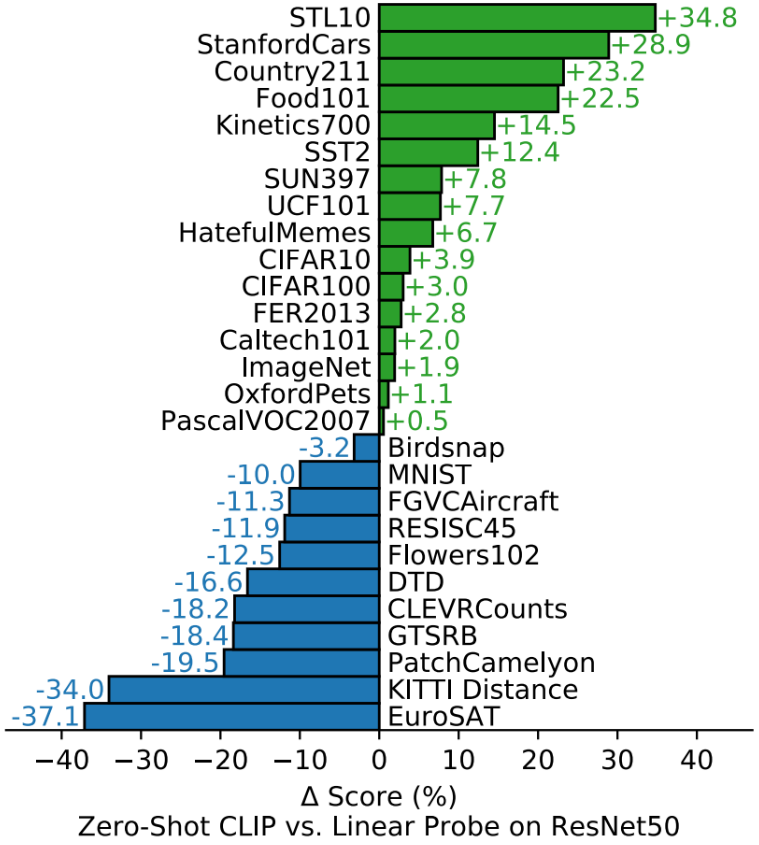
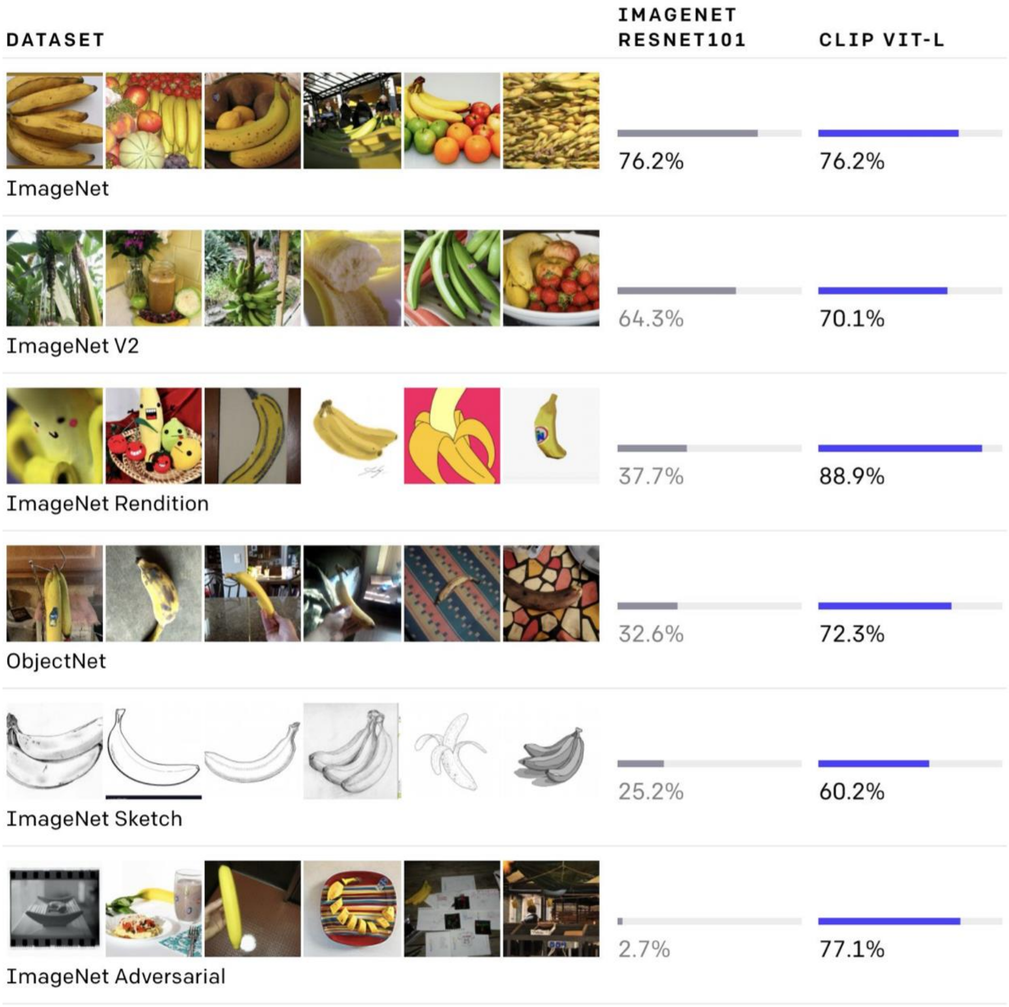


*Figure 5.* **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

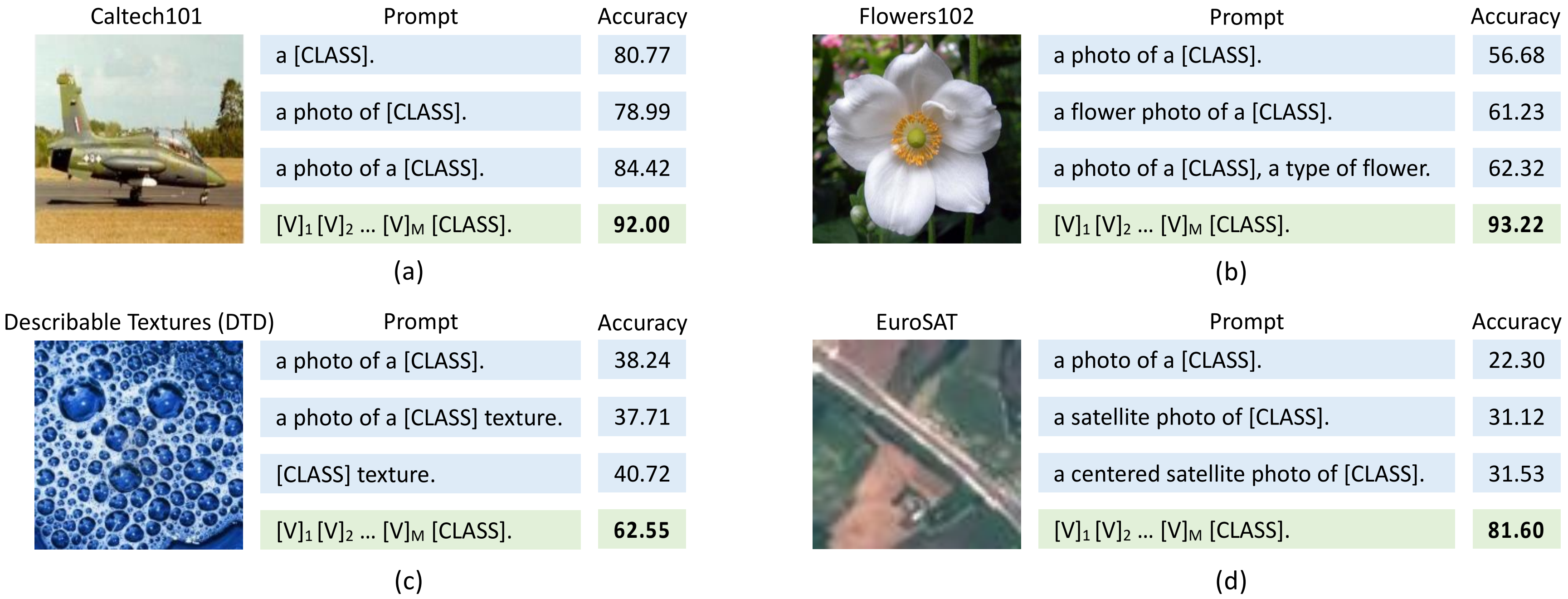


| DATASET | IMAGENET RESNET101 | CLIP VIT-L |
|---|---|---|
| ImageNet | 76.2% | 76.2% |
| ImageNet V2 | 64.3% | 70.1% |
| ImageNet Rendition | 37.7% | 88.9% |
| ObjectNet | 32.6% | 72.3% |
| ImageNet Sketch | 25.2% | 60.2% |
| ImageNet Adversarial | 2.7% | 77.1% |

Radford et al. Learning transferable visual models from natural language supervision. ICML 2021.

# Problem with hand-crafted prompt

- Difficult to tune the context words



| Caltech101 | Prompt | Accuracy |
|---|---|---|
| | a [CLASS]. | 80.77 |
| | a photo of [CLASS]. | 78.99 |
| | a photo of a [CLASS]. | 84.42 |
| | $[V]_1 [V]_2 \ldots [V]_M$ [CLASS]. | **92.00** |

(a)

| Flowers102 | Prompt | Accuracy |
|---|---|---|
| | a photo of a [CLASS]. | 56.68 |
| | a flower photo of a [CLASS]. | 61.23 |
| | a photo of a [CLASS], a type of flower. | 62.32 |
| | $[V]_1 [V]_2 \ldots [V]_M$ [CLASS]. | **93.22** |

(b)

| Describable Textures (DTD) | Prompt | Accuracy |
|---|---|---|
| | a photo of a [CLASS]. | 38.24 |
| | a photo of a [CLASS] texture. | 37.71 |
| | [CLASS] texture. | 40.72 |
| | $[V]_1 [V]_2 \ldots [V]_M$ [CLASS]. | **62.55** |

(c)

| EuroSAT | Prompt | Accuracy |
|---|---|---|
| | a photo of a [CLASS]. | 22.30 |
| | a satellite photo of [CLASS]. | 31.12 |
| | a centered satellite photo of [CLASS]. | 31.53 |
| | $[V]_1 [V]_2 \ldots [V]_M$ [CLASS]. | **81.60** |

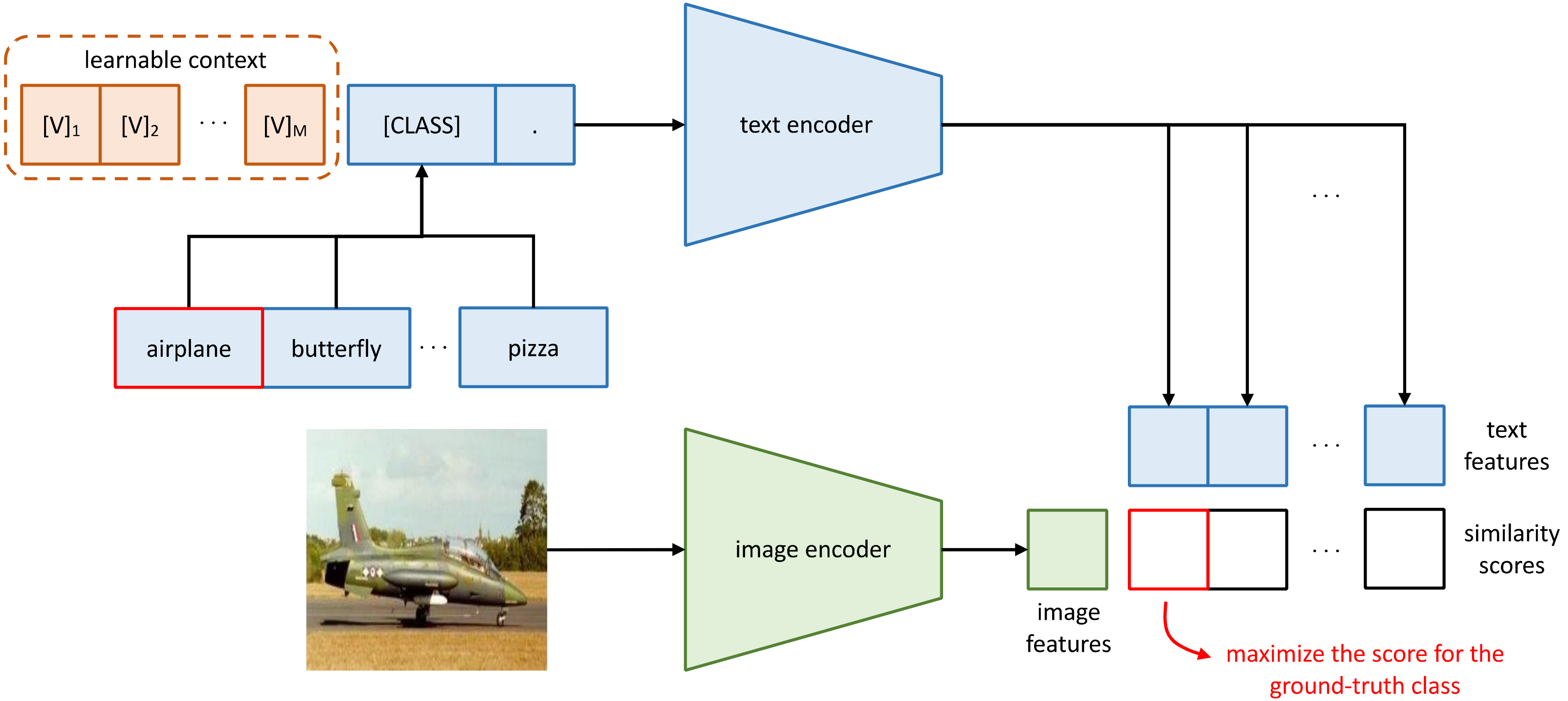
(d)

Question: Can we instead learn the context? (Yes, use prompt learning!)

Zhou et al. Learning to prompt for vision-language models. arXiv preprint 2021.

# Context optimization (CoOp)

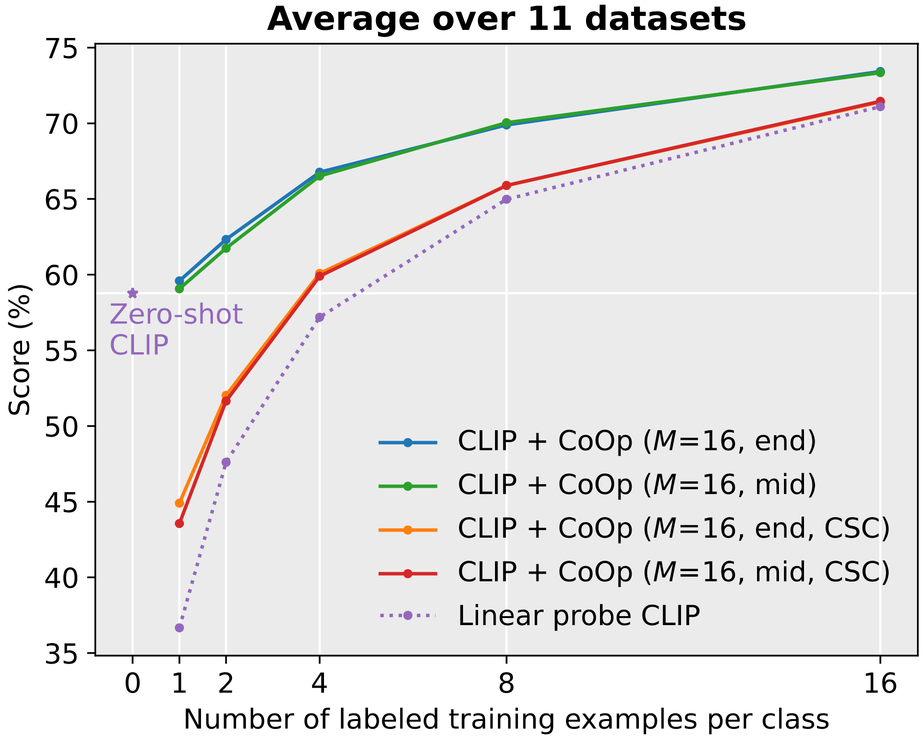
- Main idea: turn the context words into learnable vectors

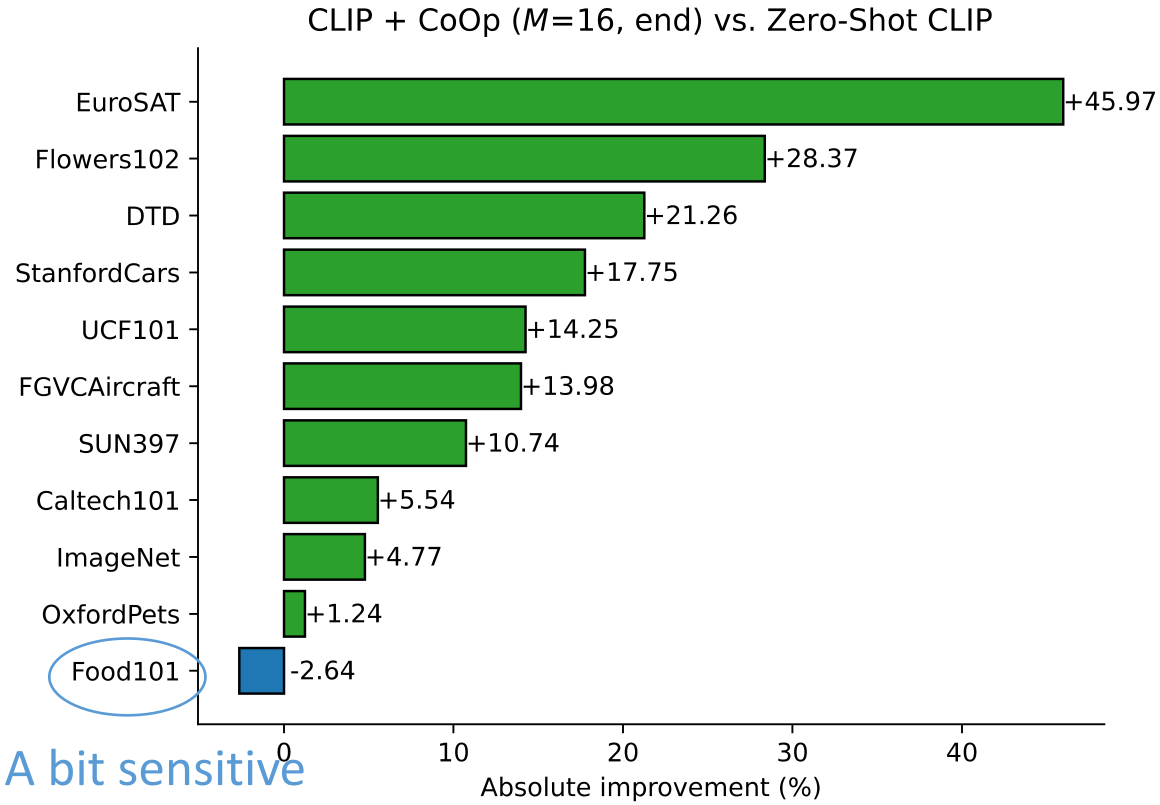


Zhou et al. Learning to prompt for vision-language models. arXiv preprint 2021.

# Pros: CoOp is a few-shot learner

- Evaluation on 11 datasets: ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT and UCF101





A bit sensitive to label noise

Zhou et al. Learning to prompt for vision-language models. arXiv preprint 2021.

# Pros: CoOp is robust to domain shift

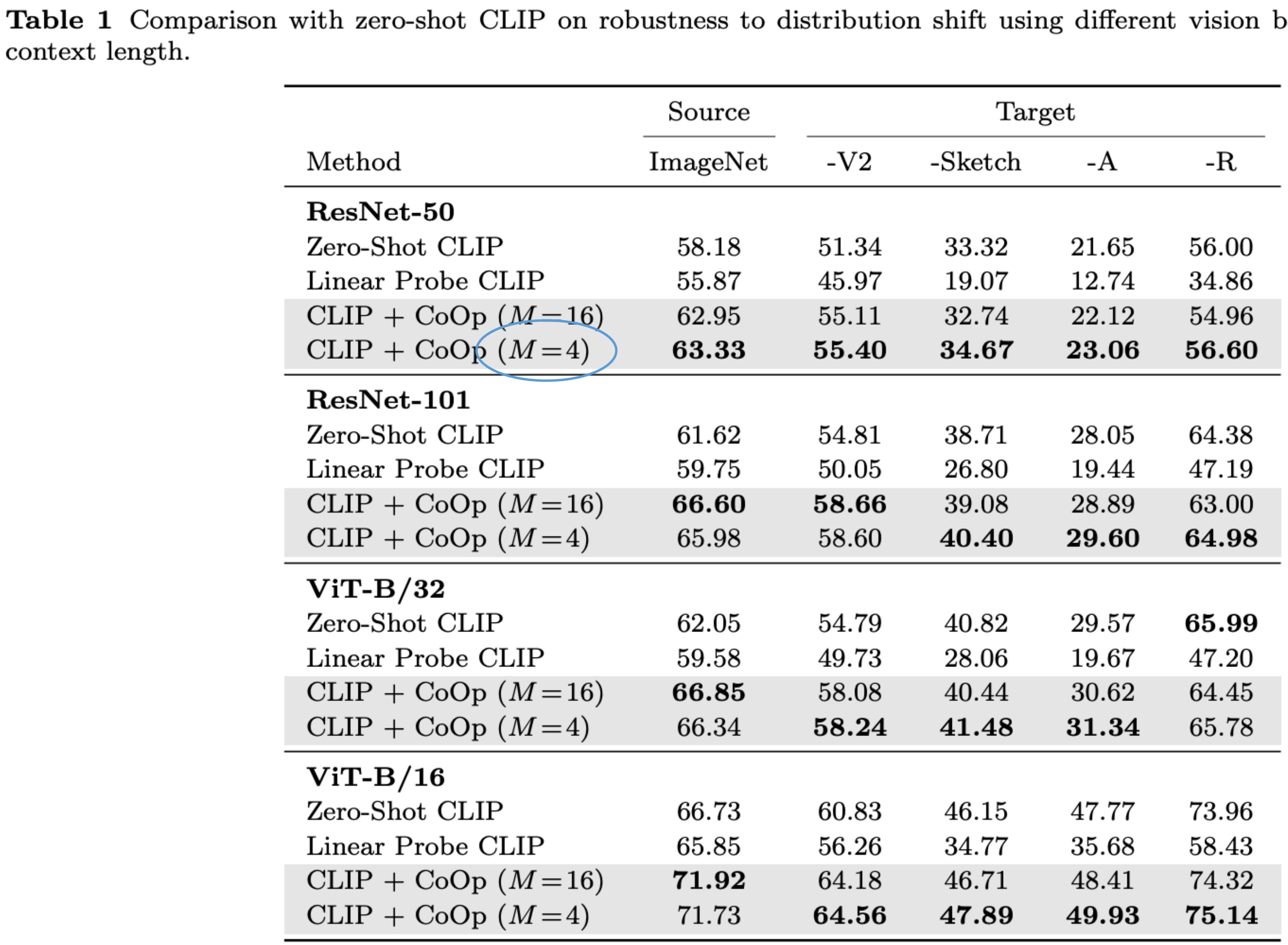


**Table 1** Comparison with zero-shot CLIP on robustness to distribution shift using different vision backbones. $M$: CoOp's context length.

| | Source | Target | | | |
|---|---|---|---|---|---|
| Method | ImageNet | -V2 | -Sketch | -A | -R |
| **ResNet-50** | | | | | |
| Zero-Shot CLIP | 58.18 | 51.34 | 33.32 | 21.65 | 56.00 |
| Linear Probe CLIP | 55.87 | 45.97 | 19.07 | 12.74 | 34.86 |
| CLIP + CoOp ($M=16$) | 62.95 | 55.11 | 32.74 | 22.12 | 54.96 |
| CLIP + CoOp ($M=4$) | **63.33** | **55.40** | **34.67** | **23.06** | **56.60** |
| **ResNet-101** | | | | | |
| Zero-Shot CLIP | 61.62 | 54.81 | 38.71 | 28.05 | 64.38 |
| Linear Probe CLIP | 59.75 | 50.05 | 26.80 | 19.44 | 47.19 |
| CLIP + CoOp ($M=16$) | **66.60** | **58.66** | 39.08 | 28.89 | 63.00 |
| CLIP + CoOp ($M=4$) | 65.98 | 58.60 | **40.40** | **29.60** | **64.98** |
| **ViT-B/32** | | | | | |
| Zero-Shot CLIP | 62.05 | 54.79 | 40.82 | 29.57 | **65.99** |
| Linear Probe CLIP | 59.58 | 49.73 | 28.06 | 19.67 | 47.20 |
| CLIP + CoOp ($M=16$) | **66.85** | 58.08 | 40.44 | 30.62 | 64.45 |
| CLIP + CoOp ($M=4$) | 66.34 | **58.24** | **41.48** | **31.34** | 65.78 |
| **ViT-B/16** | | | | | |
| Zero-Shot CLIP | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 |
| Linear Probe CLIP | 65.85 | 56.26 | 34.77 | 35.68 | 58.43 |
| CLIP + CoOp ($M=16$) | **71.92** | 64.18 | 46.71 | 48.41 | 74.32 |
| CLIP + CoOp ($M=4$) | 71.73 | **64.56** | **47.89** | **49.93** | **75.14** |

Shorter context length, better robustness

Zhou et al. Learning to prompt for vision-language models. arXiv preprint 2021.

# Cons: soft prompt learning is difficult to interpret

Conclusion: cannot use nearest words for interpretation

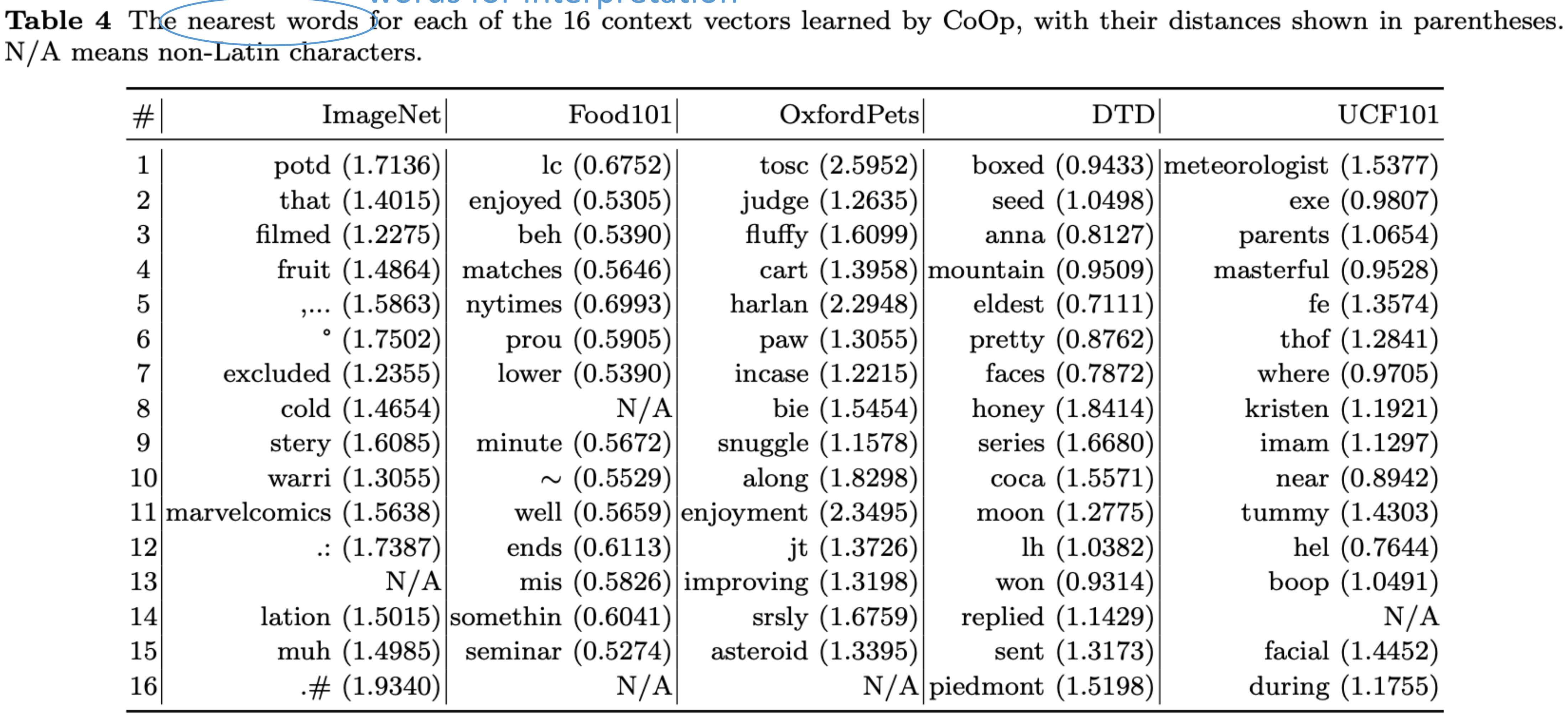


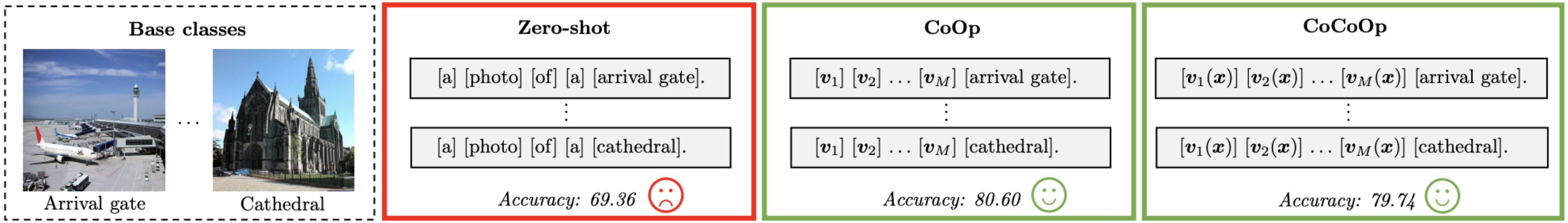
**Table 4** The nearest words for each of the 16 context vectors learned by CoOp, with their distances shown in parentheses. N/A means non-Latin characters.

| # | ImageNet | Food101 | OxfordPets | DTD | UCF101 |
|---|---|---|---|---|---|
| 1 | potd (1.7136) | lc (0.6752) | tosc (2.5952) | boxed (0.9433) | meteorologist (1.5377) |
| 2 | that (1.4015) | enjoyed (0.5305) | judge (1.2635) | seed (1.0498) | exe (0.9807) |
| 3 | filmed (1.2275) | beh (0.5390) | fluffy (1.6099) | anna (0.8127) | parents (1.0654) |
| 4 | fruit (1.4864) | matches (0.5646) | cart (1.3958) | mountain (0.9509) | masterful (0.9528) |
| 5 | ,… (1.5863) | nytimes (0.6993) | harlan (2.2948) | eldest (0.7111) | fe (1.3574) |
| 6 | ° (1.7502) | prou (0.5905) | paw (1.3055) | pretty (0.8762) | thof (1.2841) |
| 7 | excluded (1.2355) | lower (0.5390) | incase (1.2215) | faces (0.7872) | where (0.9705) |
| 8 | cold (1.4654) | N/A | bie (1.5454) | honey (1.8414) | kristen (1.1921) |
| 9 | stery (1.6085) | minute (0.5672) | snuggle (1.1578) | series (1.6680) | imam (1.1297) |
| 10 | warri (1.3055) | ∼ (0.5529) | along (1.8298) | coca (1.5571) | near (0.8942) |
| 11 | marvelcomics (1.5638) | well (0.5659) | enjoyment (2.3495) | moon (1.2775) | tummy (1.4303) |
| 12 | .: (1.7387) | ends (0.6113) | jt (1.3726) | lh (1.0382) | hel (0.7644) |
| 13 | N/A | mis (0.5826) | improving (1.3198) | won (0.9314) | boop (1.0491) |
| 14 | lation (1.5015) | somethin (0.6041) | srsly (1.6759) | replied (1.1429) | N/A |
| 15 | muh (1.4985) | seminar (0.5274) | asteroid (1.3395) | sent (1.3173) | facial (1.4452) |
| 16 | .# (1.9340) | N/A | N/A | piedmont (1.5198) | during (1.1755) |

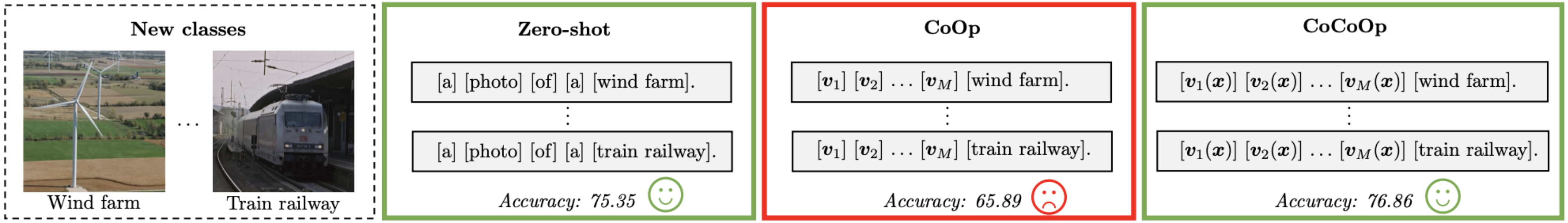
Zhou et al. Learning to prompt for vision-language models. arXiv preprint 2021.

# Problem with CoOp

- Overfit base classes and fail to generalize to new classes



(a) Both CoOp and CoCoOp work well on the base classes observed during training and beat manual prompts by a significant margin.

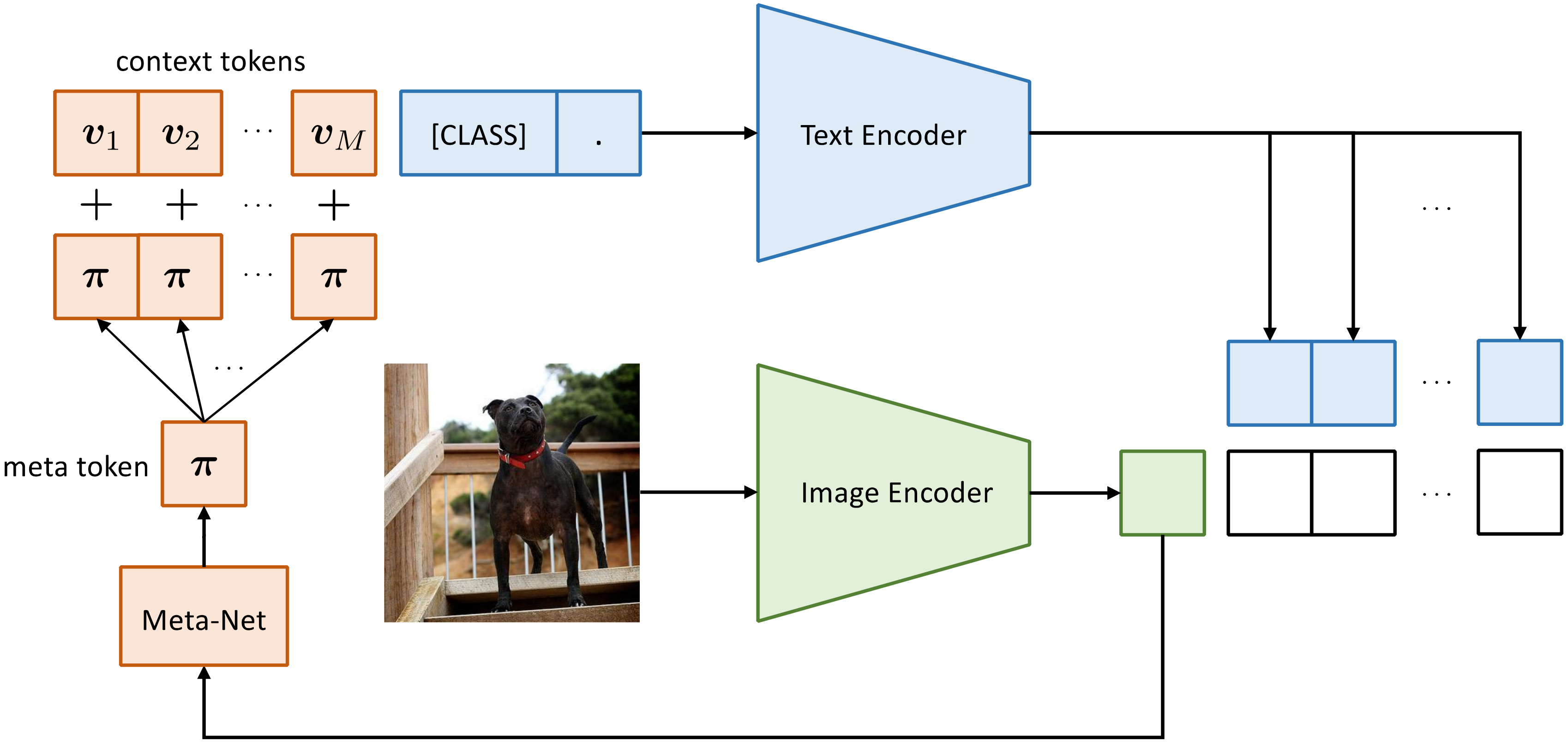


(b) The instance-conditional prompts learned by CoCoOp are much more generalizable than CoOp to the unseen classes.

Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022.

# Conditional context optimization (CoCoOp)

- Main idea: condition the context on each input image



Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022.

# Findings

- Conditional prompt learning is more generalizable

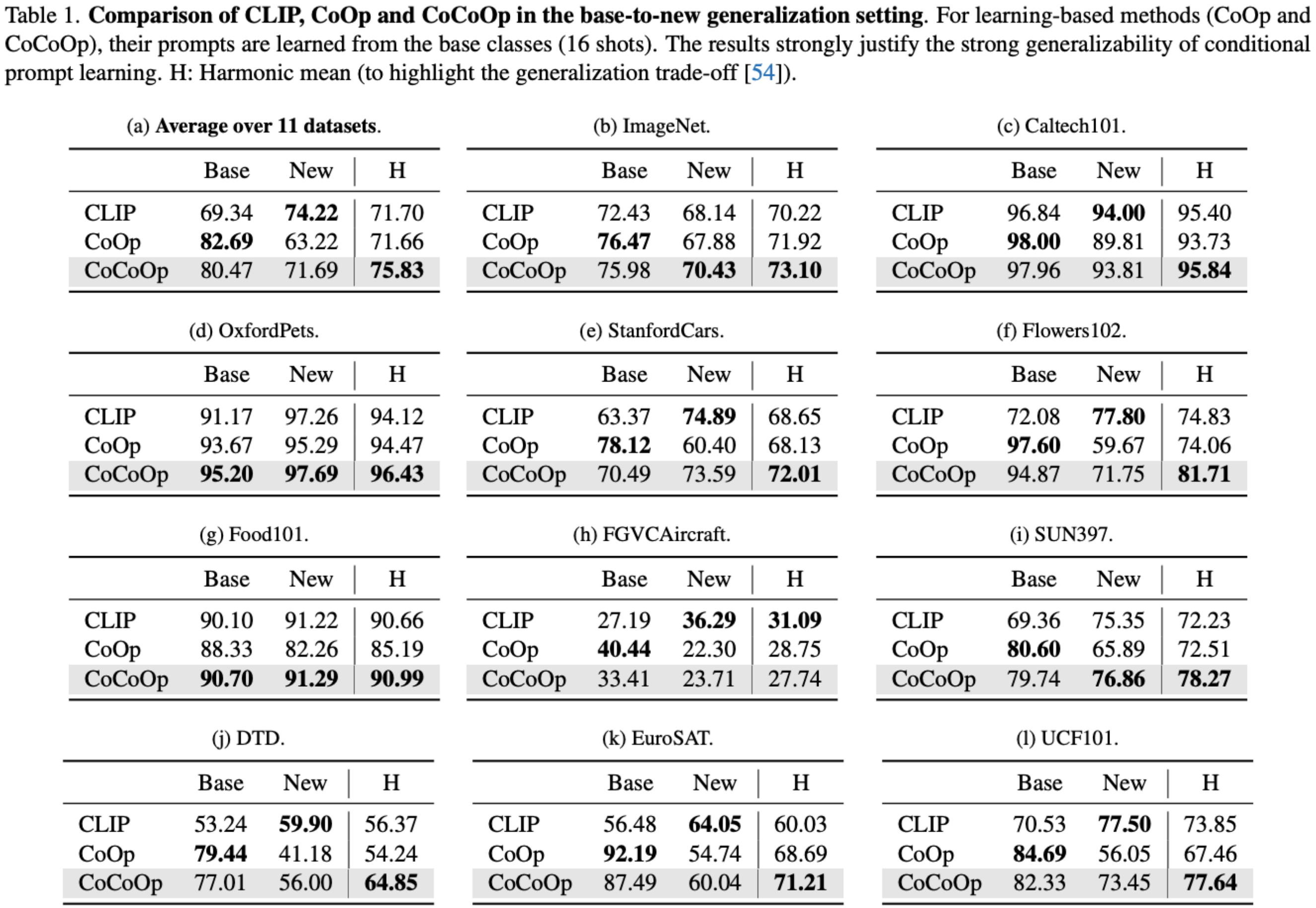


Table 1. **Comparison of CLIP, CoOp and CoCoOp in the base-to-new generalization setting**. For learning-based methods (CoOp and CoCoOp), their prompts are learned from the base classes (16 shots). The results strongly justify the strong generalizability of conditional prompt learning. H: Harmonic mean (to highlight the generalization trade-off [54]).

(a) **Average over 11 datasets**.

| | Base | New | H |
|---|---|---|---|
| CLIP | 69.34 | **74.22** | 71.70 |
| CoOp | **82.69** | 63.22 | 71.66 |
| CoCoOp | 80.47 | 71.69 | **75.83** |

(b) ImageNet.

| | Base | New | H |
|---|---|---|---|
| CLIP | 72.43 | 68.14 | 70.22 |
| CoOp | **76.47** | 67.88 | 71.92 |
| CoCoOp | 75.98 | **70.43** | **73.10** |

(c) Caltech101.

| | Base | New | H |
|---|---|---|---|
| CLIP | 96.84 | **94.00** | 95.40 |
| CoOp | **98.00** | 89.81 | 93.73 |
| CoCoOp | 97.96 | 93.81 | **95.84** |

(d) OxfordPets.

| | Base | New | H |
|---|---|---|---|
| CLIP | 91.17 | 97.26 | 94.12 |
| CoOp | 93.67 | 95.29 | 94.47 |
| CoCoOp | **95.20** | **97.69** | **96.43** |

(e) StanfordCars.

| | Base | New | H |
|---|---|---|---|
| CLIP | 63.37 | **74.89** | 68.65 |
| CoOp | **78.12** | 60.40 | 68.13 |
| CoCoOp | 70.49 | 73.59 | **72.01** |

(f) Flowers102.

| | Base | New | H |
|---|---|---|---|
| CLIP | 72.08 | **77.80** | 74.83 |
| CoOp | **97.60** | 59.67 | 74.06 |
| CoCoOp | 94.87 | 71.75 | **81.71** |

(g) Food101.

| | Base | New | H |
|---|---|---|---|
| CLIP | 90.10 | 91.22 | 90.66 |
| CoOp | 88.33 | 82.26 | 85.19 |
| CoCoOp | **90.70** | **91.29** | **90.99** |

(h) FGVCAircraft.

| | Base | New | H |
|---|---|---|---|
| CLIP | 27.19 | **36.29** | **31.09** |
| CoOp | **40.44** | 22.30 | 28.75 |
| CoCoOp | 33.41 | 23.71 | 27.74 |

(i) SUN397.

| | Base | New | H |
|---|---|---|---|
| CLIP | 69.36 | 75.35 | 72.23 |
| CoOp | **80.60** | 65.89 | 72.51 |
| CoCoOp | 79.74 | **76.86** | **78.27** |

(j) DTD.

| | Base | New | H |
|---|---|---|---|
| CLIP | 53.24 | **59.90** | 56.37 |
| CoOp | **79.44** | 41.18 | 54.24 |
| CoCoOp | 77.01 | 56.00 | **64.85** |

(k) EuroSAT.

| | Base | New | H |
|---|---|---|---|
| CLIP | 56.48 | **64.05** | 60.03 |
| CoOp | **92.19** | 54.74 | 68.69 |
| CoCoOp | 87.49 | 60.04 | **71.21** |

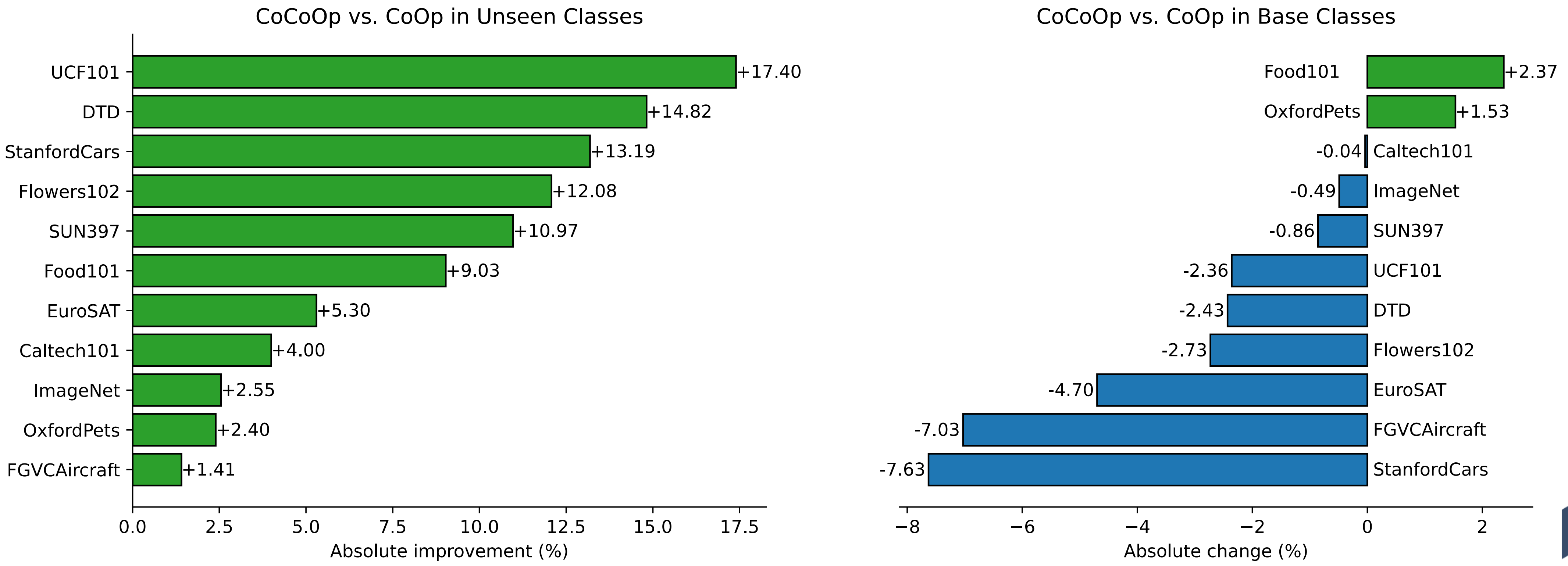
(l) UCF101.

| | Base | New | H |
|---|---|---|---|
| CLIP | 70.53 | **77.50** | 73.85 |
| CoOp | **84.69** | 56.05 | 67.46 |
| CoCoOp | 82.33 | 73.45 | **77.64** |

Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022.

# Findings

- Sacrifice accuracy on base classes but the gains on generalization are larger



(a) (b)

Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022.

# Findings

- Conditional prompt learning is also more transferable

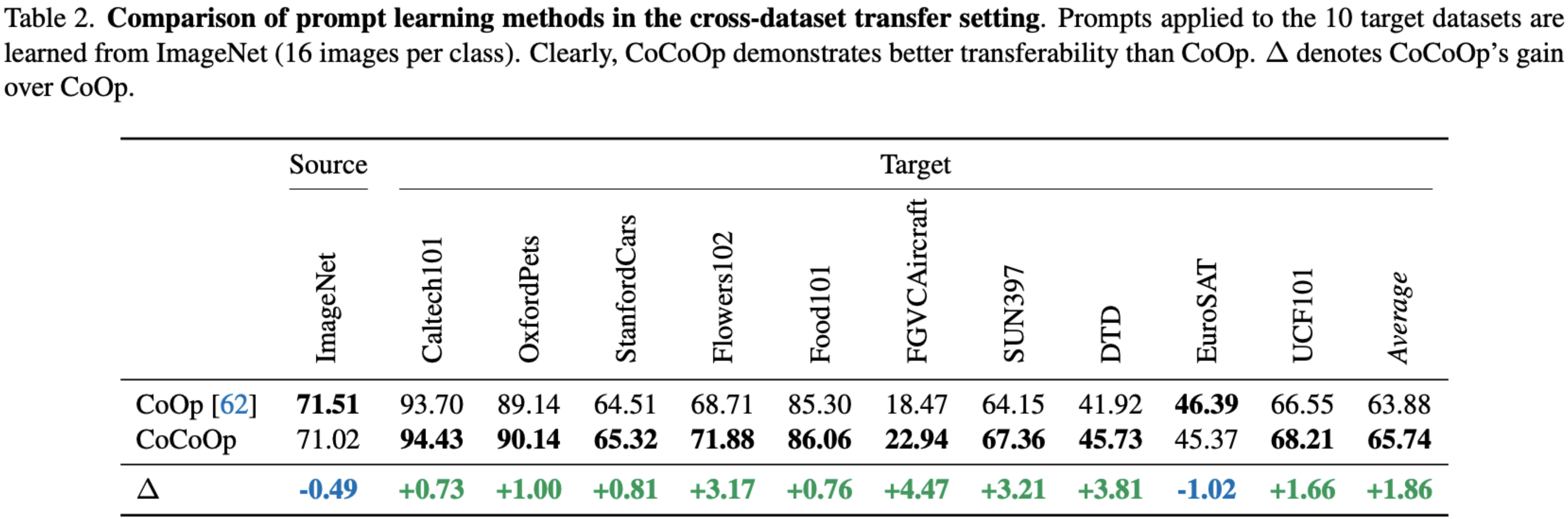


Table 2. **Comparison of prompt learning methods in the cross-dataset transfer setting**. Prompts applied to the 10 target datasets are learned from ImageNet (16 images per class). Clearly, CoCoOp demonstrates better transferability than CoOp. Δ denotes CoCoOp's gain over CoOp.

| | Source | Target | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ImageNet | Caltech101 | OxfordPets | StanfordCars | Flowers102 | Food101 | FGVCAircraft | SUN397 | DTD | EuroSAT | UCF101 | *Average* |
| CoOp [62] | **71.51** | 93.70 | 89.14 | 64.51 | 68.71 | 85.30 | 18.47 | 64.15 | 41.92 | **46.39** | 66.55 | 63.88 |
| CoCoOp | 71.02 | **94.43** | **90.14** | **65.32** | **71.88** | **86.06** | **22.94** | **67.36** | **45.73** | 45.37 | **68.21** | **65.74** |
| Δ | -0.49 | +0.73 | +1.00 | +0.81 | +3.17 | +0.76 | +4.47 | +3.21 | +3.81 | -1.02 | +1.66 | +1.86 |

Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022.

# Findings

- More robust to domain shift as well

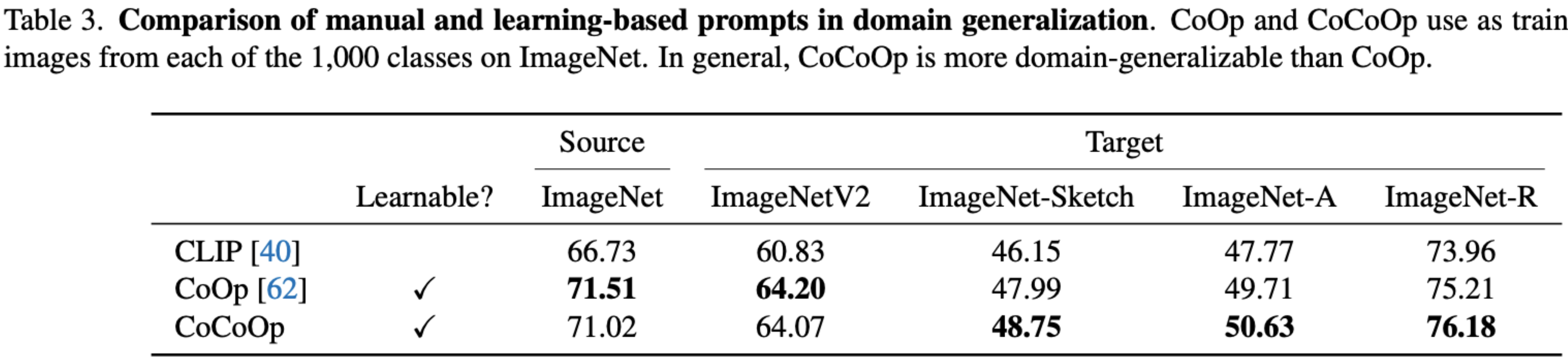


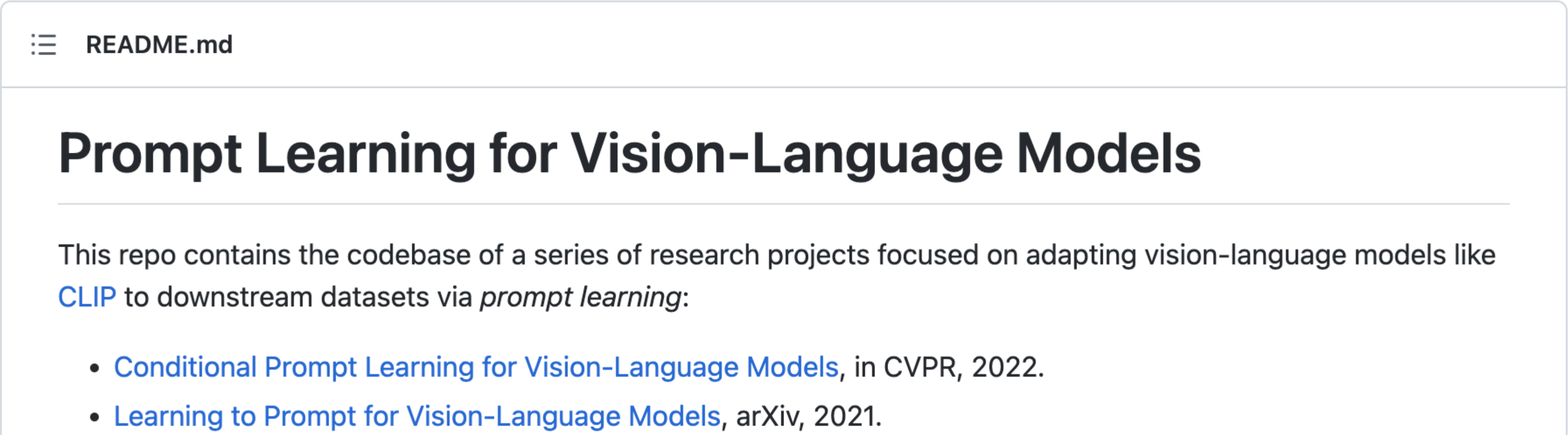
Table 3. **Comparison of manual and learning-based prompts in domain generalization**. CoOp and CoCoOp use as training data 16 images from each of the 1,000 classes on ImageNet. In general, CoCoOp is more domain-generalizable than CoOp.

| | | Source | Target | | | |
|---|---|---|---|---|---|---|
| | Learnable? | ImageNet | ImageNetV2 | ImageNet-Sketch | ImageNet-A | ImageNet-R |
| CLIP [40] | | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 |
| CoOp [62] | ✓ | **71.51** | **64.20** | 47.99 | 49.71 | 75.21 |
| CoCoOp | ✓ | 71.02 | 64.07 | **48.75** | **50.63** | **76.18** |

Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022.

# Code and models

- Released at [https://github.com/KaiyangZhou/CoOp](https://github.com/KaiyangZhou/CoOp)



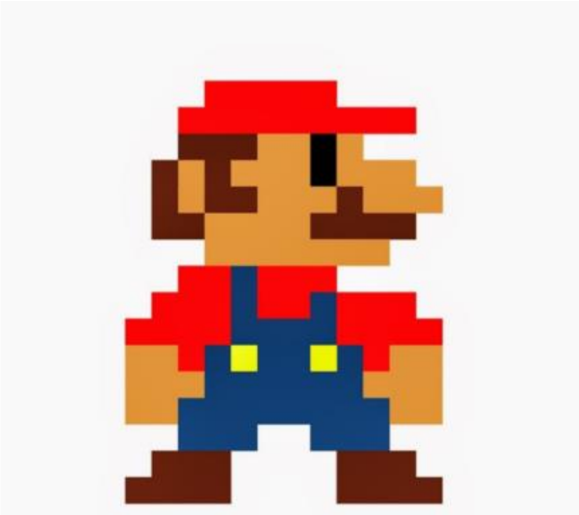
README.md

## Prompt Learning for Vision-Language Models

This repo contains the codebase of a series of research projects focused on adapting vision-language models like CLIP to downstream datasets via *prompt learning*:

- Conditional Prompt Learning for Vision-Language Models, in CVPR, 2022.
- Learning to Prompt for Vision-Language Models, arXiv, 2021.

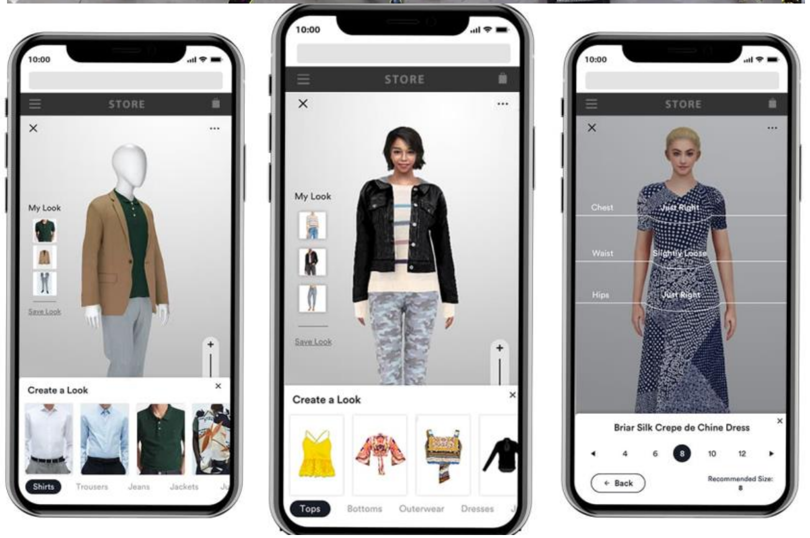
# 2D + 3D



Hong et al., Versatile Multi-Modal Pre-Training for Human-Centric Perception, CVPR 2022 (Oral)

# Why Human-Centric Pre-train?

## Vital role in many applications

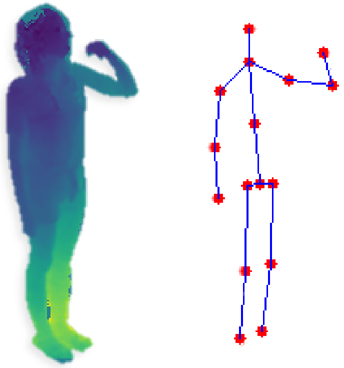


## Expensive and dense annotations

DensePose

Part Segmentation

Part Segmentation



3D Keypoints

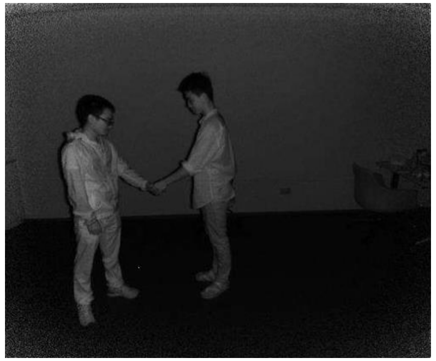
# Multi-modal Nature of Human Data

## Dense representations

*Pros: rich texture/ 3D geometry*
*Cons: low-level and noisy*

RGB

Depth

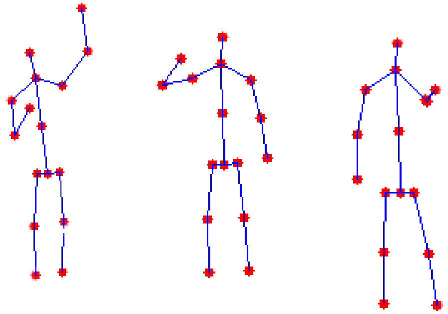


Infrared

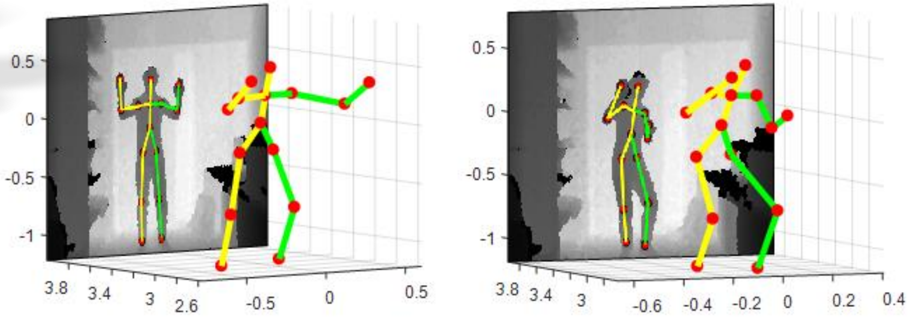
How to combine both in pre-train?

## Sparse representations

*Pros: rich in semantics and structured*
*Cons: insufficient details*

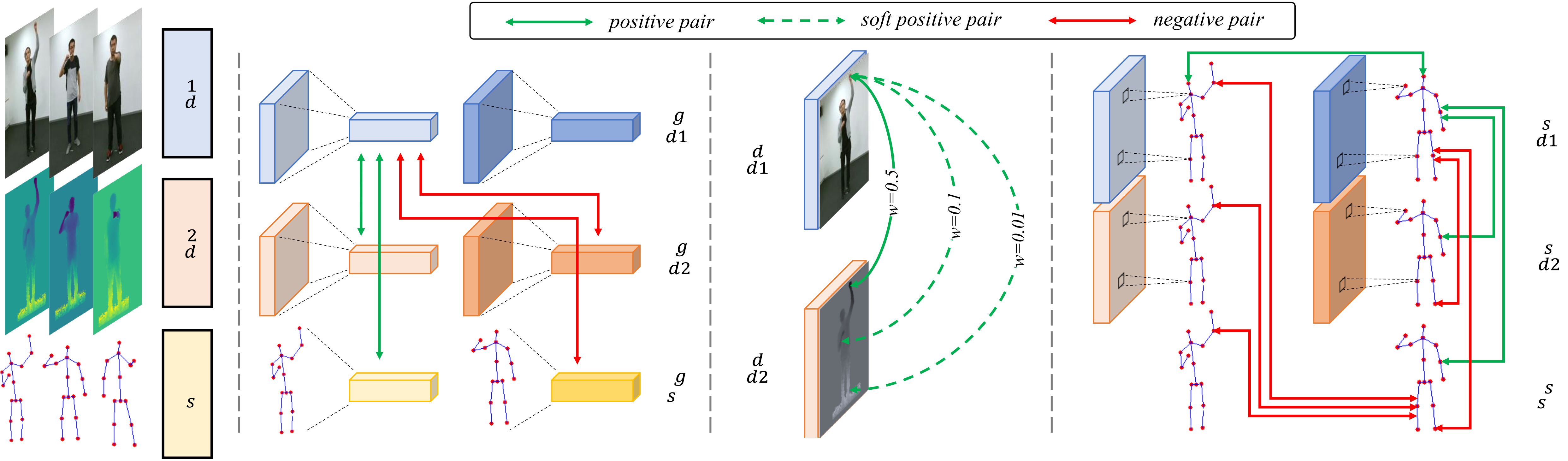


2D Keypoints



3D Keypoints

# HCMoCo – Principles of Learning Targets



*Global*

1) Mutual Information Maximization

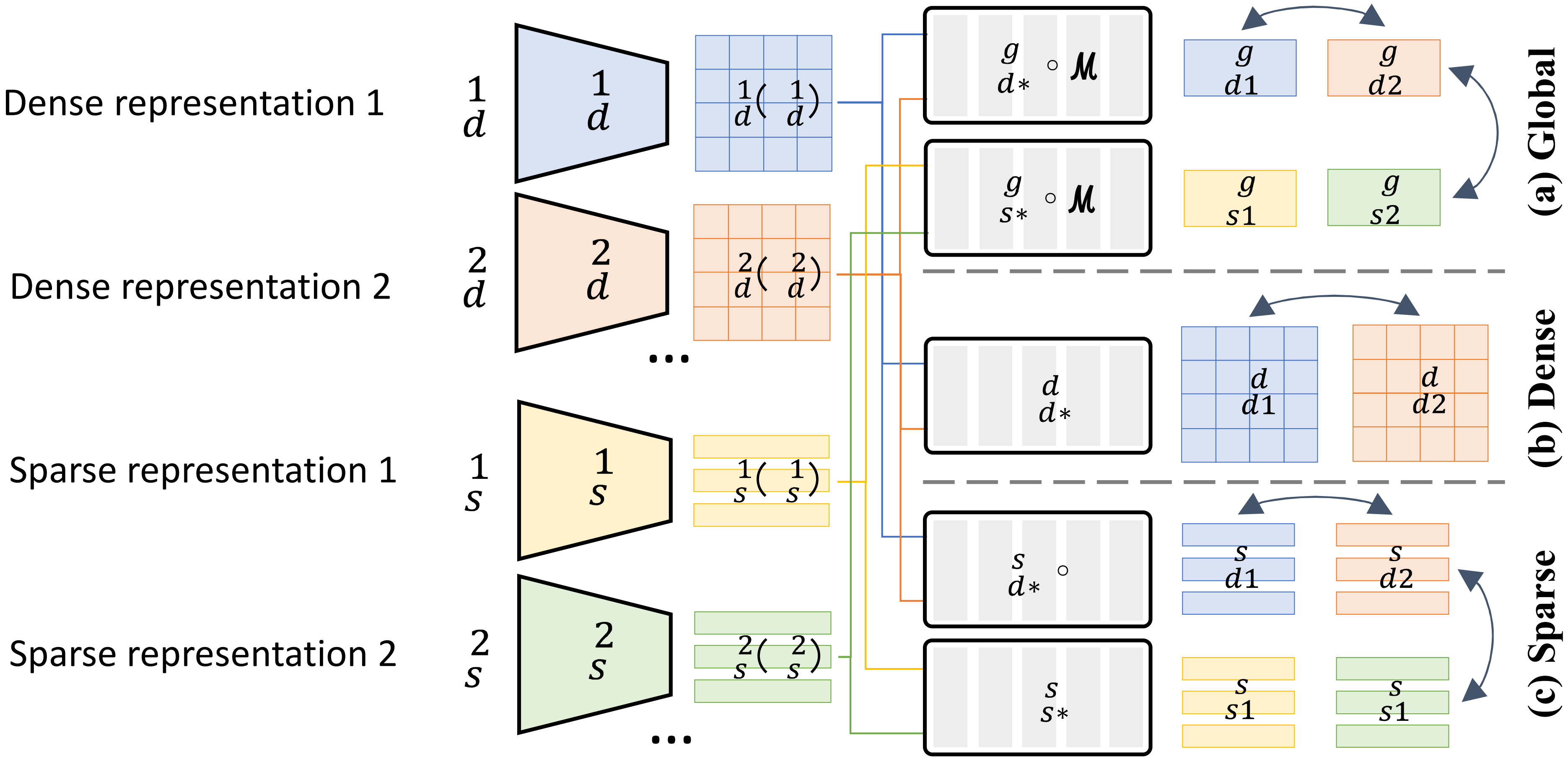
*Dense*

2) Continuous and Ordinal Feature Distribution

*Sparse*

3) Structure-Aware Semantic Consistency

# HCMoCo – General Paradigm



Dense representation 1

Dense representation 2

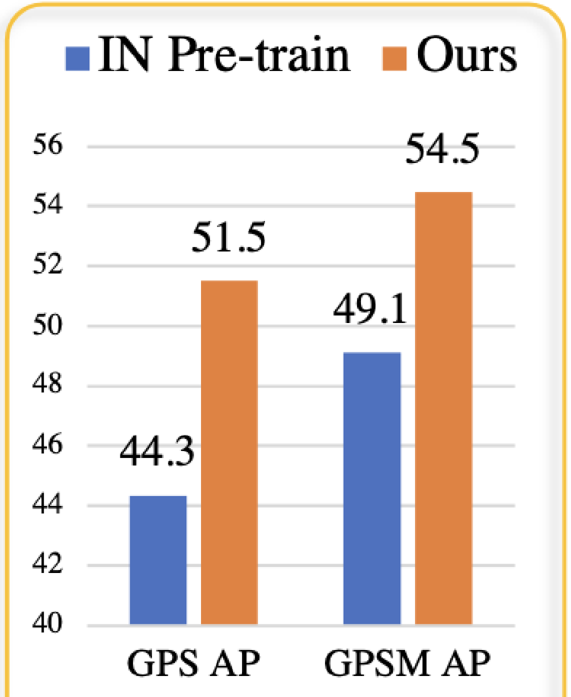
Sparse representation 1

Sparse representation 2

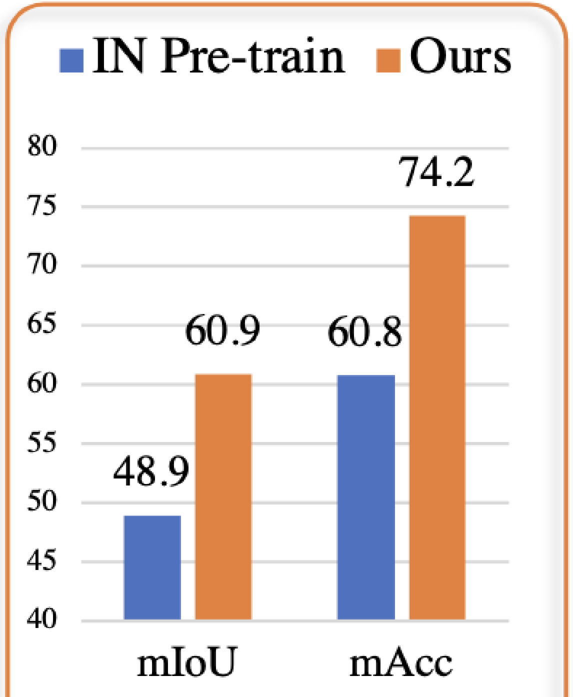
*Hierarchical Contrastive Learning Targets*

# High Performance on Downstream Tasks

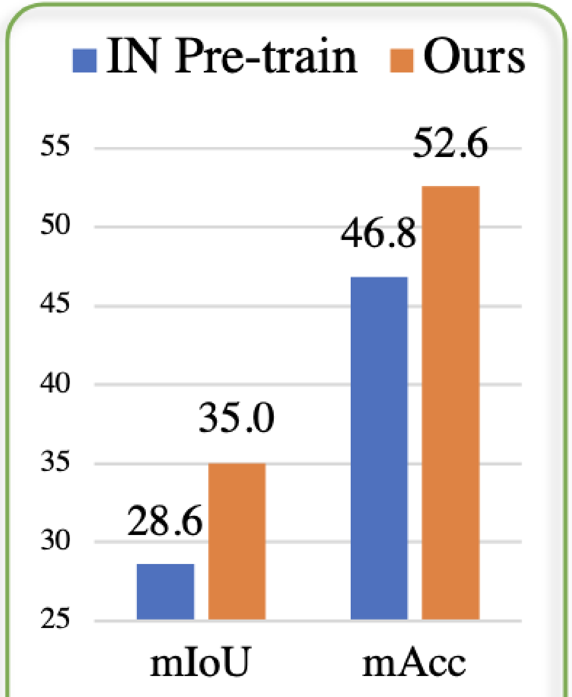
One-time pre-training, boost the performance of all the downstream tasks of multiple modalities.



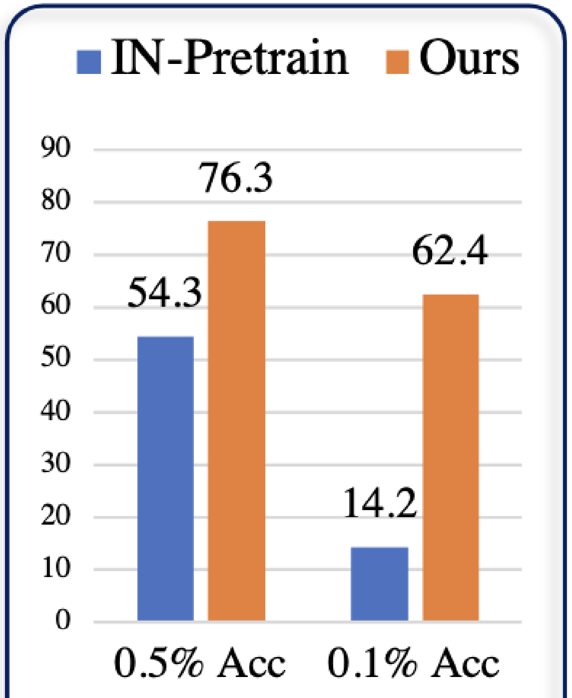
*1) DensePose Estimation*



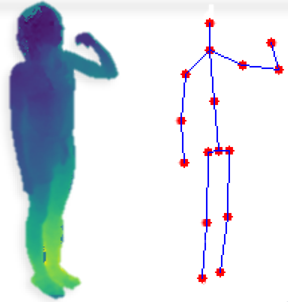
*2) RGB Human Parsing*



*3) Depth Human Parsing*

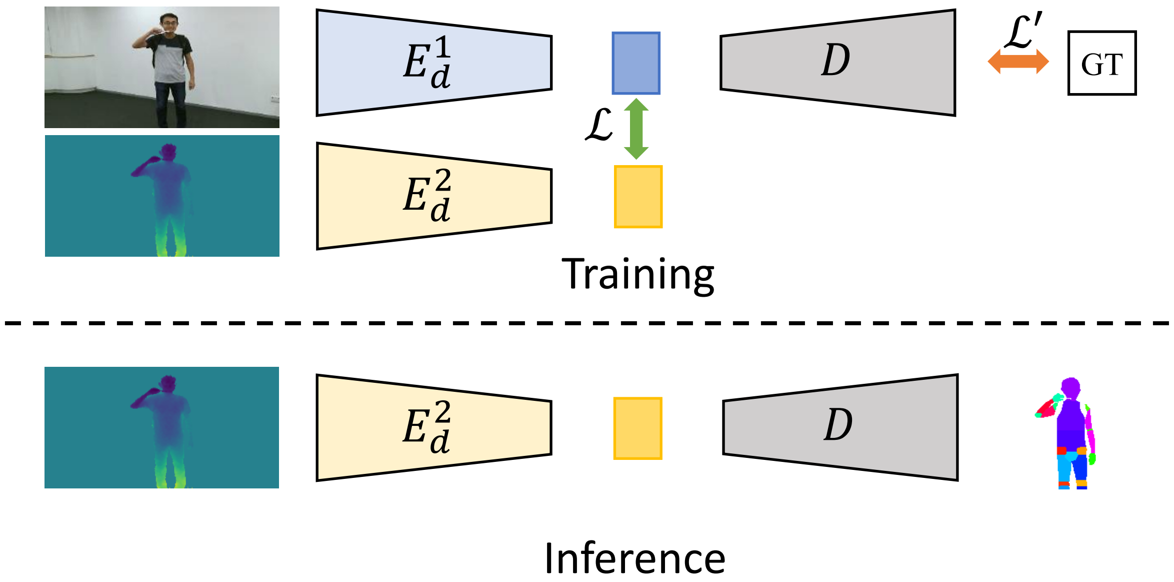


*4) Depth 3D Pose Est.*



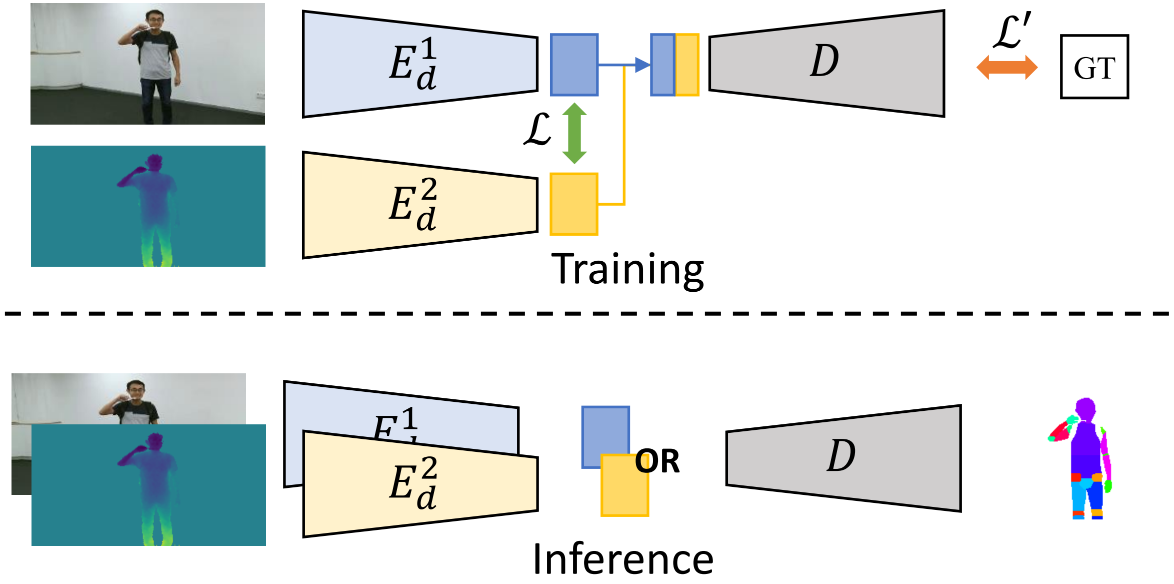
# Versatility of HCMoCo

### (a) Cross-Modality Supervision



| Method | RGB → Depth | | | Depth → RGB | | |
|---|---|---|---|---|---|---|
| | mIoU | mAcc | aAcc | mIoU | mAcc | aAcc |
| No Contrastive | 3.94 | 4.36 | 92.24 | 3.71 | 4.03 | 91.63 |
| CMC [44] | 3.86 | 5.59 | 86.81 | 3.85 | 4.27 | 91.75 |
| Ours | **33.19** | **54.38** | **94.70** | **26.80** | **48.80** | **92.84** |

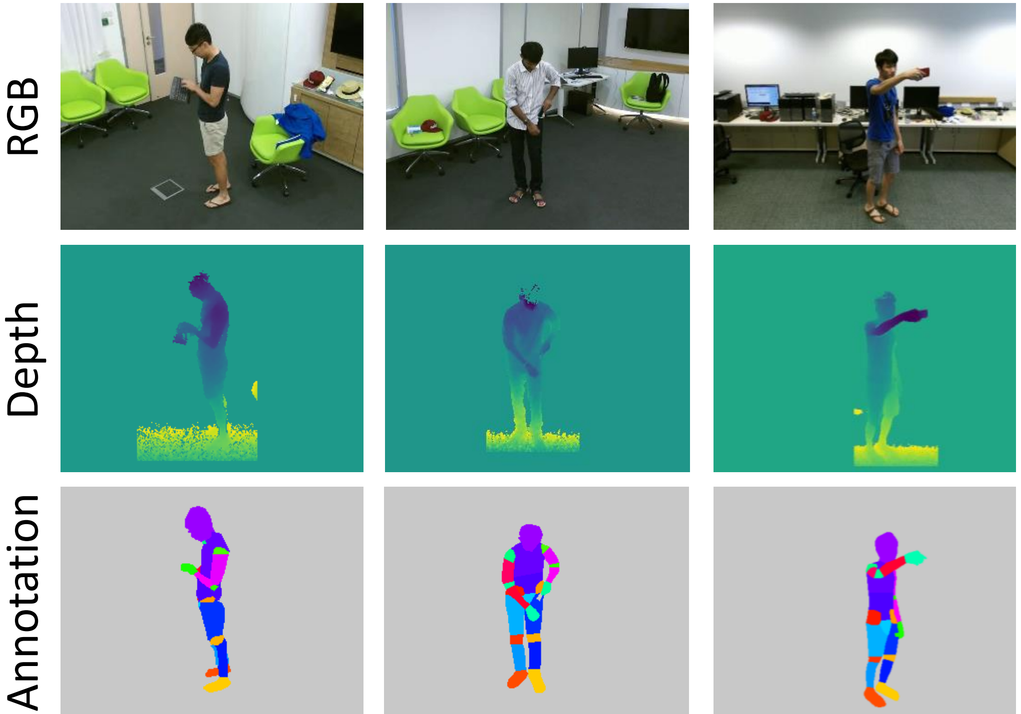
### (b) Missing-Modality Inference

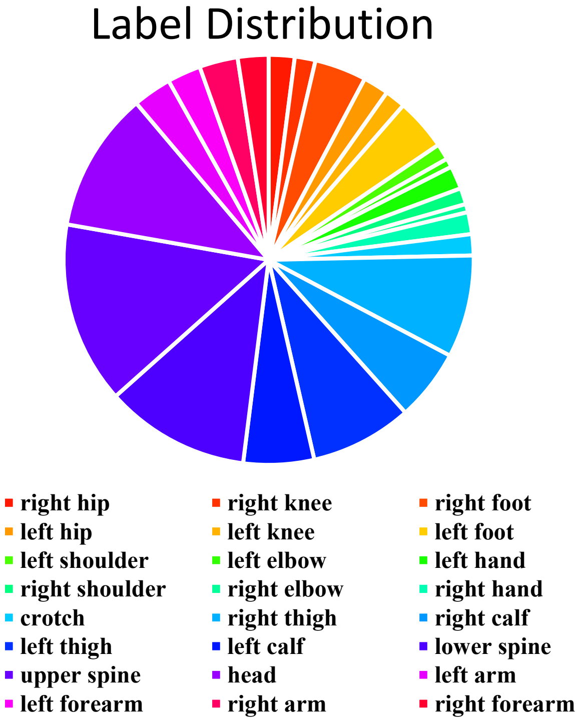


| Method | Only RGB | | | Only Depth | | |
|---|---|---|---|---|---|---|
| | mIoU | mAcc | aAcc | mIoU | mAcc | aAcc |
| No Contrastive | 13.45 | 14.77 | 93.35 | 24.41 | 30.49 | 95.27 |
| CMC [44] | 19.62 | 28.19 | 92.94 | 16.58 | 19.83 | 93.94 |
| Ours | **43.88** | **64.27** | **96.15** | **43.98** | **63.66** | **96.34** |

# Dataset – NTURGBD-Parsing-4K

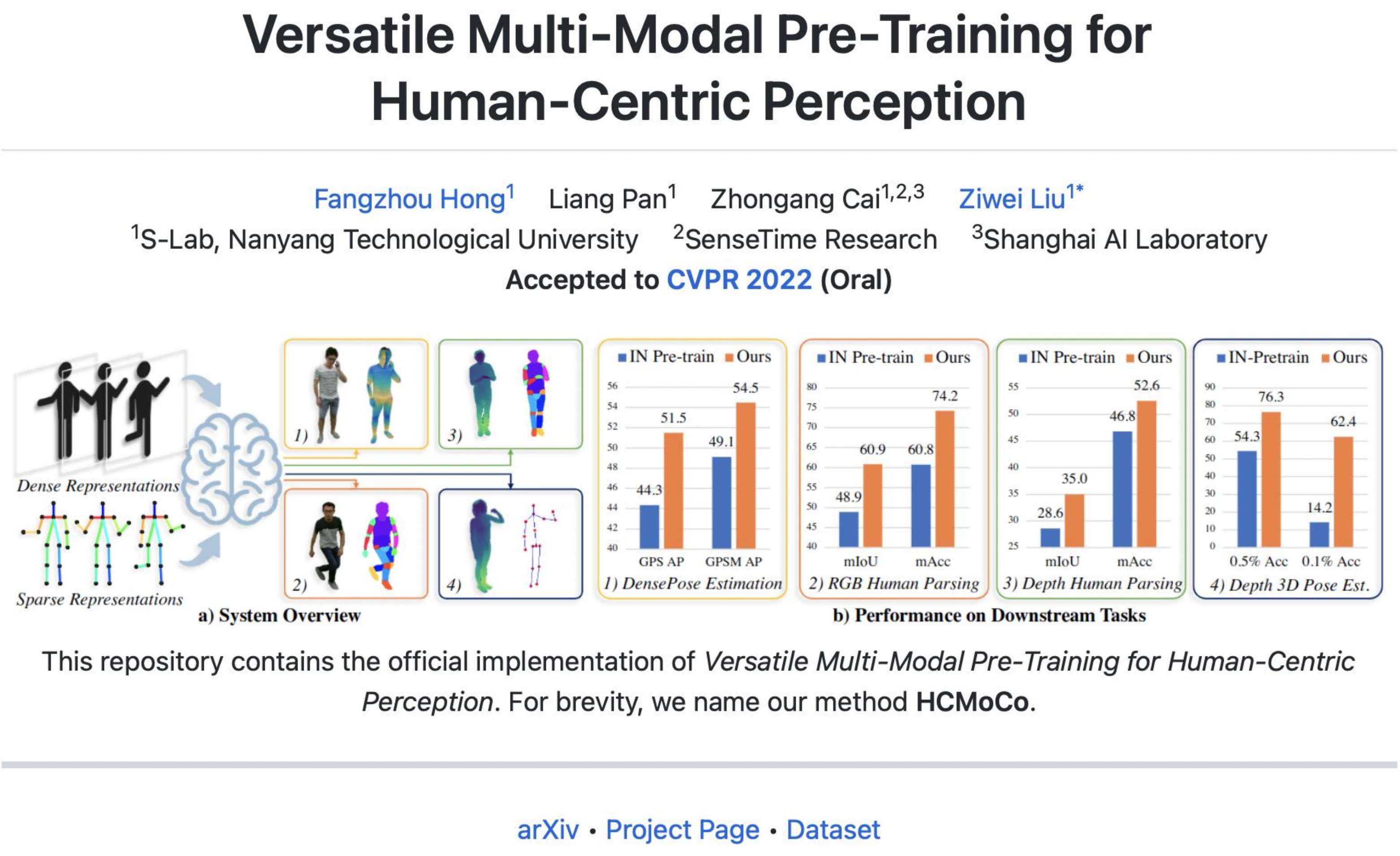
- The first RGB-D human parsing dataset
- Uniformly sampled 3926 samples from NTU RGB+D (60/120)
- Annotate 24 human body parts





# Code, Models & Dataset

Released at https://github.com/hongfz16/HCMoCo



## Versatile Multi-Modal Pre-Training for Human-Centric Perception

Fangzhou Hong[1] Liang Pan[1] Zhongang Cai[1,2,3] Ziwei Liu[1*]
[1]S-Lab, Nanyang Technological University [2]SenseTime Research [3]Shanghai AI Laboratory

**Accepted to CVPR 2022 (Oral)**

a) System Overview

b) Performance on Downstream Tasks

This repository contains the official implementation of *Versatile Multi-Modal Pre-Training for Human-Centric Perception*. For brevity, we name our method **HCMoCo**.

arXiv • Project Page • Dataset

# Generalization in Vision Models

**Semantic Shift**

***OOD Detection***

*Zero-shot / Few-shot / Long-tailed Learning*

Sensory Modalities

Neural Architectures

*Corruptions / Perturbations / Domain Shifts*

Covariate Shift

# Out-of-Distribution Detection

Yang et al., Generalized Out-of-Distribution Detection: A Survey, ArXiv 2021
Yang et al., Full-Spectrum Out-of-Distribution Detection, ArXiv 2022

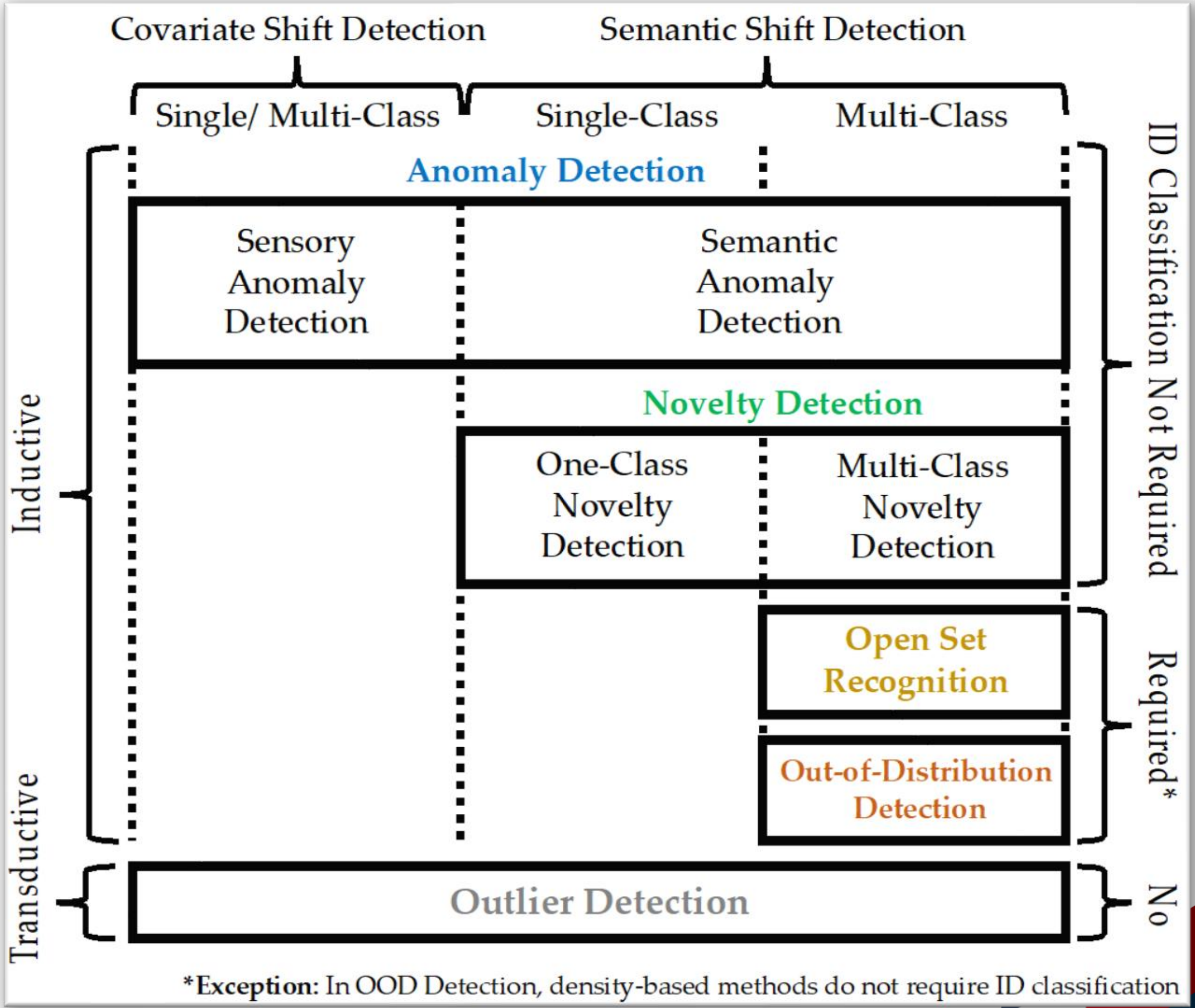
# Generalized OOD Detection: A Survey

**Why We Write The Survey:**

- Several topics share quite similar goals:
    - Anomaly Detection (AD)
    - Novelty Detection (ND)
    - Open Set Recognition (OSR)
    - Out-of-Distribution (OOD) Detection
    - Outlier Detection (OD)

- We discuss the commonalty and difference among them to eliminate the confusion for practitioners and newcomers.

- A generic framework **generalized OOD detection** is proposed to encompasses all five problems, which can be seen as special cases or sub-tasks and are easier to distinguish.

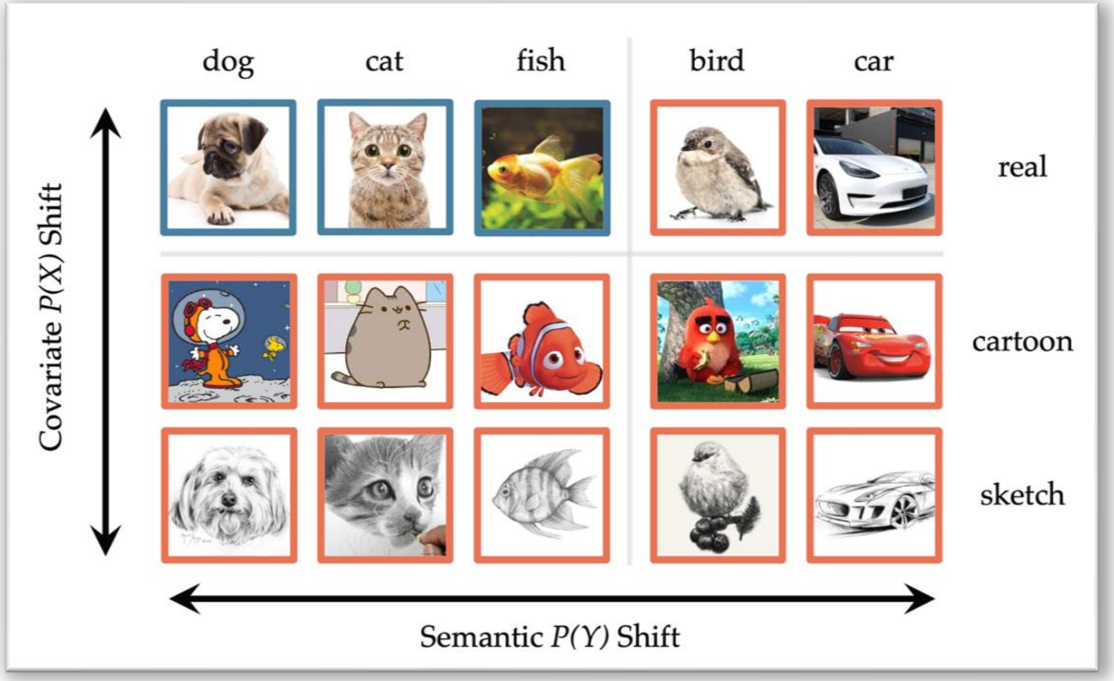


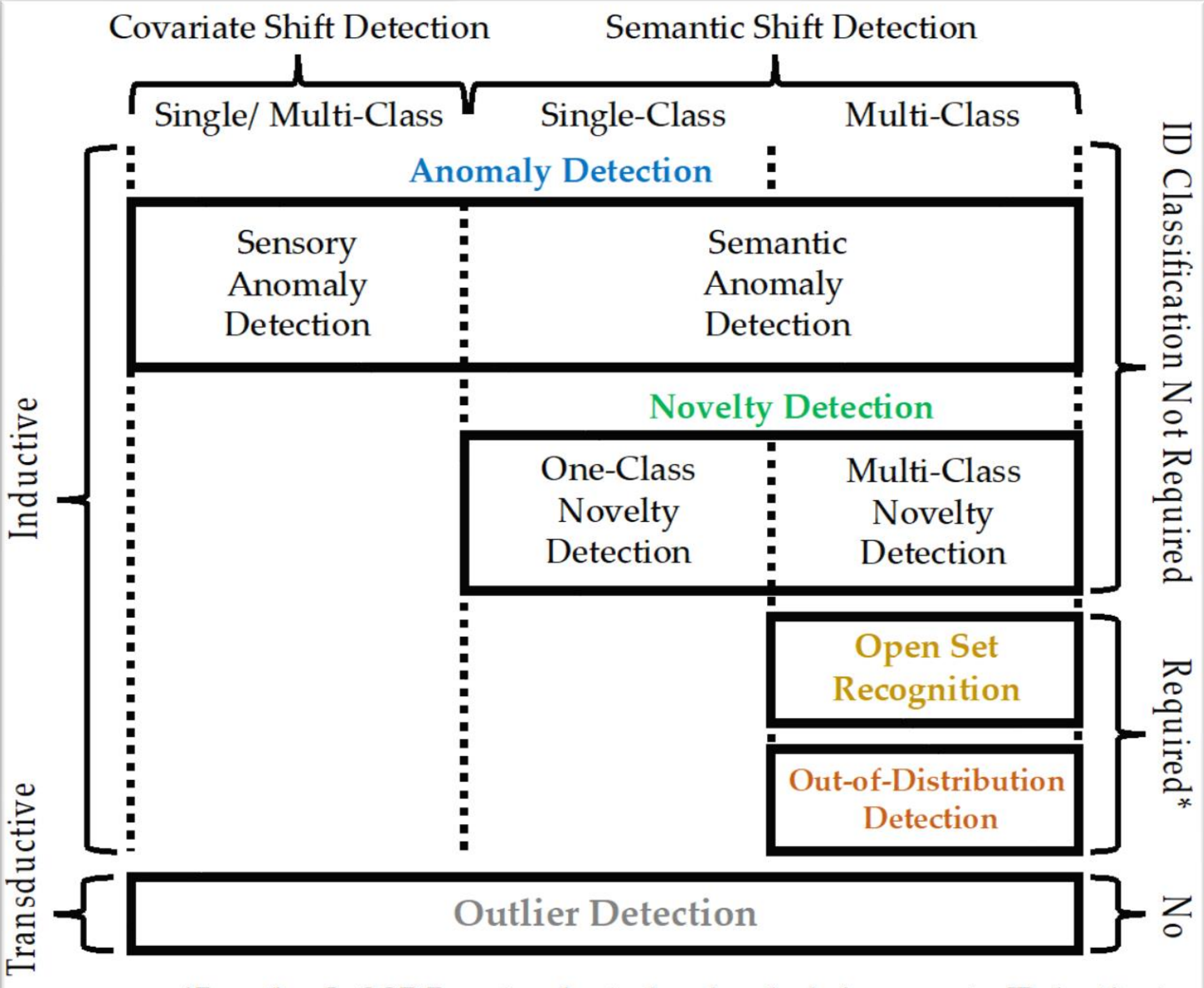
https://github.com/Jingkang50/OODSurvey

[Jingkang Yang, Kaiyang Zhou, Yixuan Li, Ziwei Liu. Generalized OOD Detection: A Survey. arXiv:2110.11334. 2021 ]

# Generalized OOD Detection: A Survey

**Generic Framework:**
**- Generalized OOD Detection**





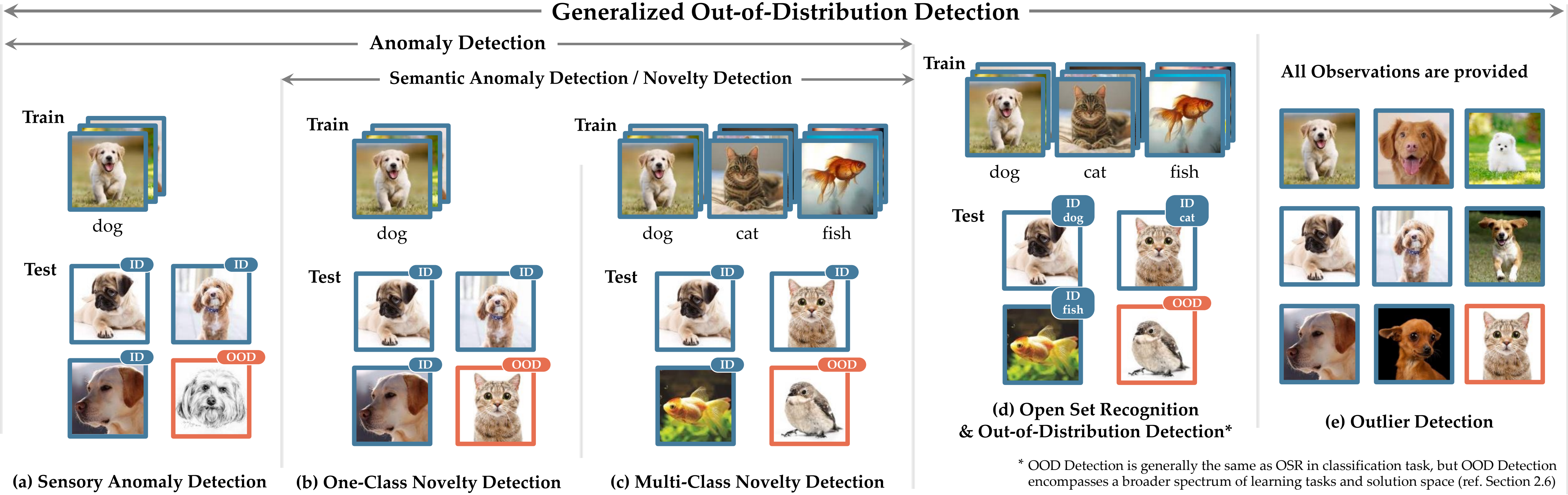
*Exception: In OOD Detection, density-based methods do not require ID classification

https://github.com/Jingkang50/OODSurvey

[Jingkang Yang, Kaiyang Zhou, Yixuan Li, Ziwei Liu. Generalized OOD Detection: A Survey. arXiv:2110.11334. 2021 ]

# Generalized OOD Detection: A Survey

**Generic Framework:**
**- Generalized OOD Detection**



(a) Sensory Anomaly Detection (b) One-Class Novelty Detection (c) Multi-Class Novelty Detection (d) Open Set Recognition & Out-of-Distribution Detection* (e) Outlier Detection

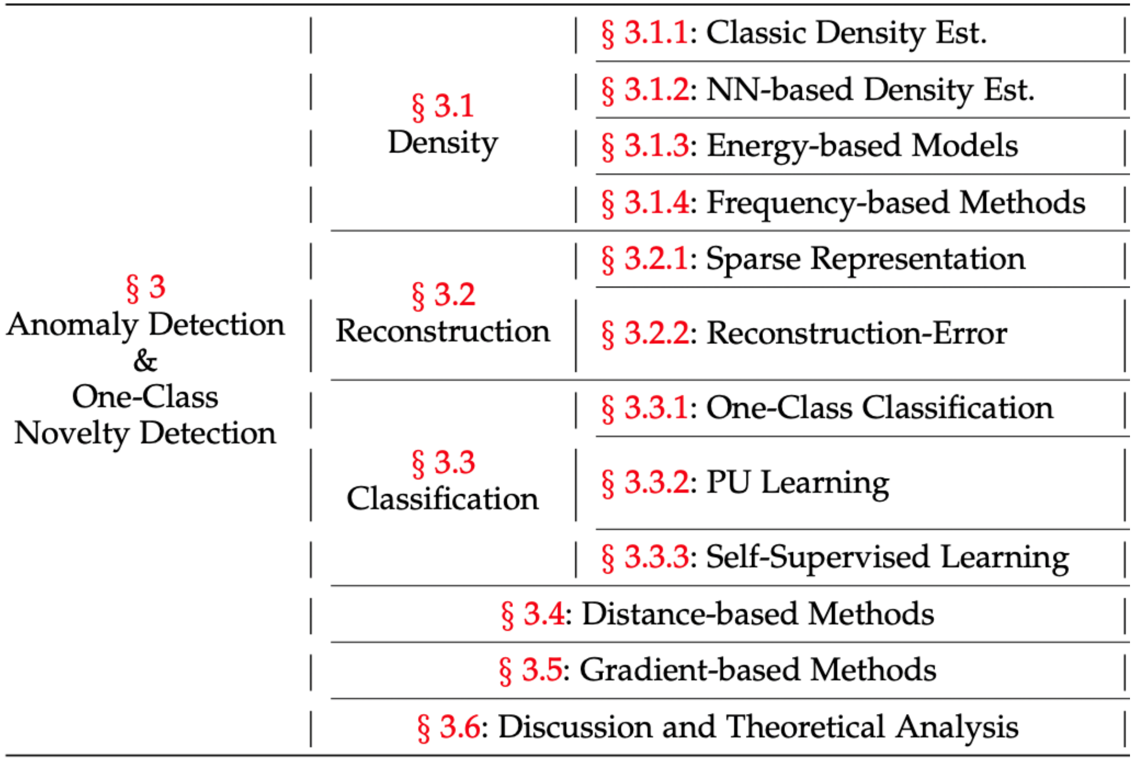
* OOD Detection is generally the same as OSR in classification task, but OOD Detection encompasses a broader spectrum of learning tasks and solution space (ref. Section 2.6)

https://github.com/Jingkang50/OODSurvey

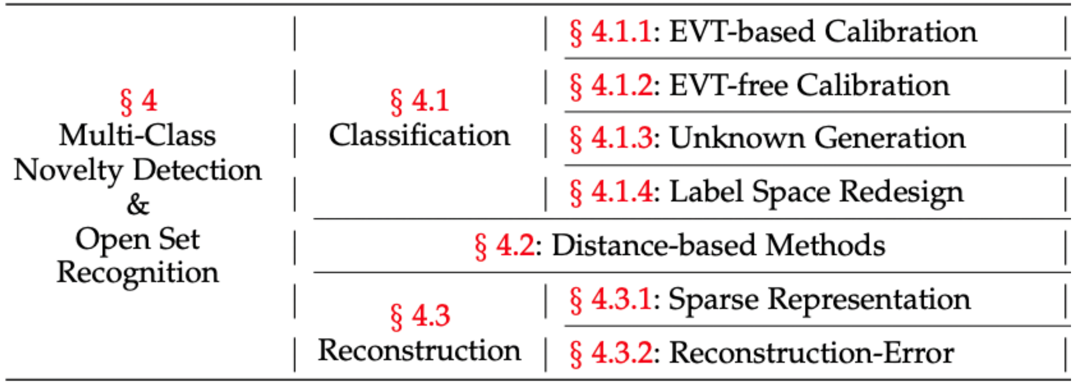
[Jingkang Yang, Kaiyang Zhou, Yixuan Li, Ziwei Liu. Generalized OOD Detection: A Survey. arXiv:2110.11334. 2021 ]

# Generalized OOD Detection: A Survey

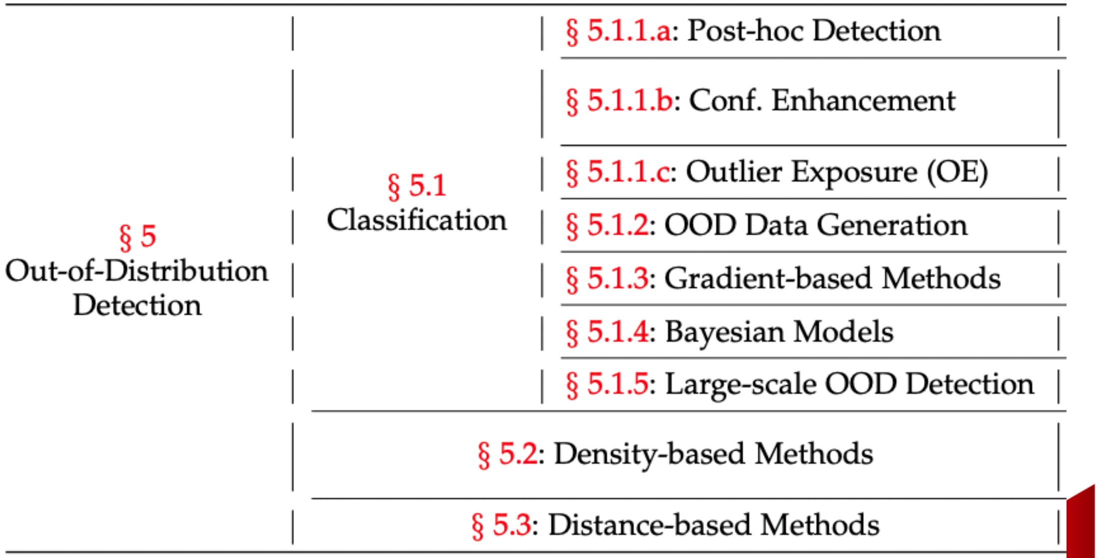
**Methodology Taxonomy**



| | | |
|---|---|---|
| § 3<br>Anomaly Detection<br>&<br>One-Class<br>Novelty Detection | § 3.1<br>Density | § 3.1.1: Classic Density Est. |
| | | § 3.1.2: NN-based Density Est. |
| | | § 3.1.3: Energy-based Models |
| | | § 3.1.4: Frequency-based Methods |
| | § 3.2<br>Reconstruction | § 3.2.1: Sparse Representation |
| | | § 3.2.2: Reconstruction-Error |
| | § 3.3<br>Classification | § 3.3.1: One-Class Classification |
| | | § 3.3.2: PU Learning |
| | | § 3.3.3: Self-Supervised Learning |
| | § 3.4: Distance-based Methods | |
| | § 3.5: Gradient-based Methods | |
| | § 3.6: Discussion and Theoretical Analysis | |



| | | |
|---|---|---|
| § 4<br>Multi-Class<br>Novelty Detection<br>&<br>Open Set<br>Recognition | § 4.1<br>Classification | § 4.1.1: EVT-based Calibration |
| | | § 4.1.2: EVT-free Calibration |
| | | § 4.1.3: Unknown Generation |
| | | § 4.1.4: Label Space Redesign |
| | § 4.2: Distance-based Methods | |
| | § 4.3<br>Reconstruction | § 4.3.1: Sparse Representation |
| | | § 4.3.2: Reconstruction-Error |



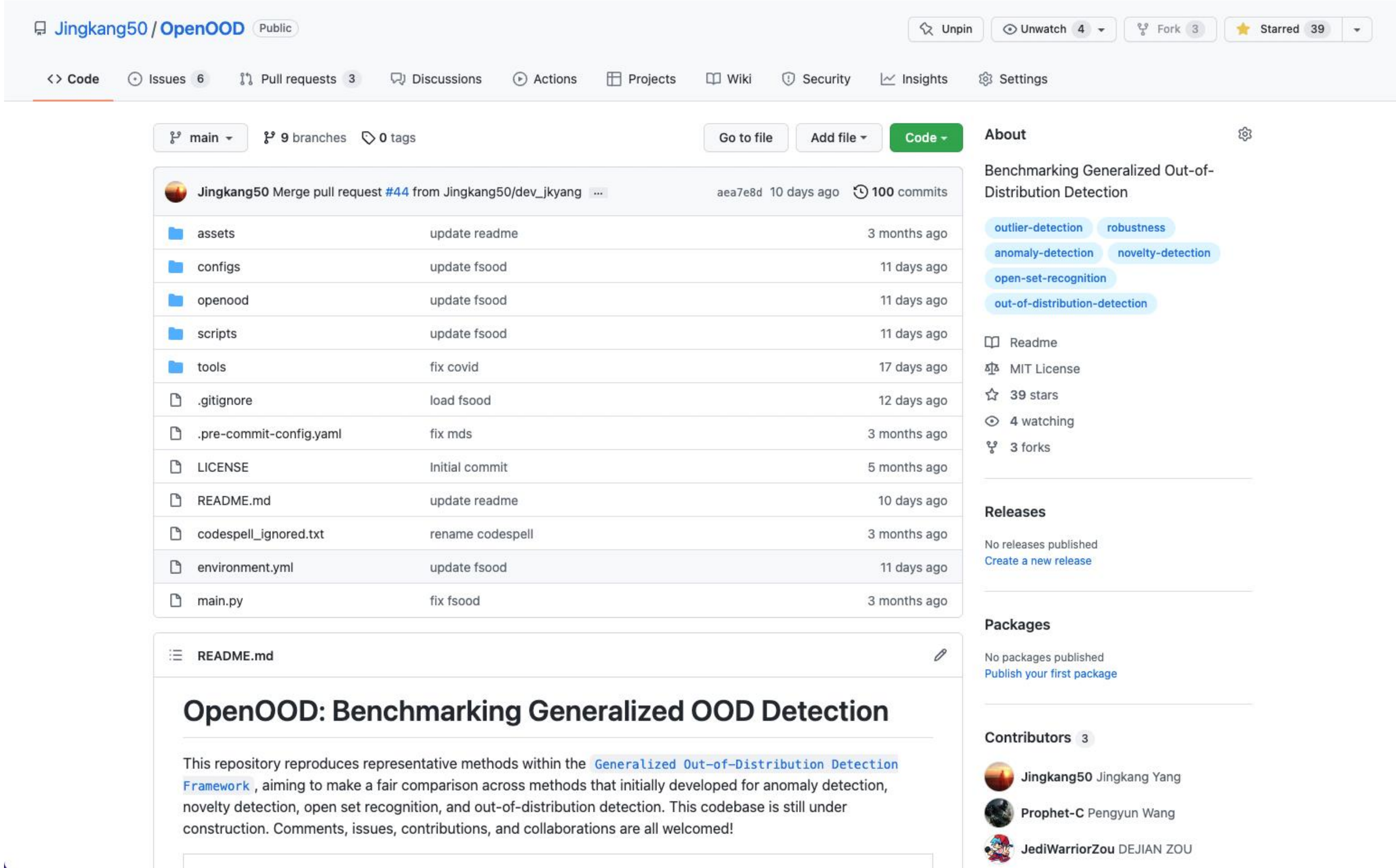
| | | |
|---|---|---|
| § 5<br>Out-of-Distribution<br>Detection | § 5.1<br>Classification | § 5.1.1.a: Post-hoc Detection |
| | | § 5.1.1.b: Conf. Enhancement |
| | | § 5.1.1.c: Outlier Exposure (OE) |
| | | § 5.1.2: OOD Data Generation |
| | | § 5.1.3: Gradient-based Methods |
| | | § 5.1.4: Bayesian Models |
| | | § 5.1.5: Large-scale OOD Detection |
| | § 5.2: Density-based Methods | |
| | § 5.3: Distance-based Methods | |

https://github.com/Jingkang50/OODSurvey

[Jingkang Yang, Kaiyang Zhou, Yixuan Li, Ziwei Liu. Generalized OOD Detection: A Survey. arXiv:2110.11334. 2021 ]

# Benchmarking Generalized OOD Detection

**OpenOOD:** https://github.com/Jingkang50/OpenOOD



Jingkang50 / OpenOOD Public

Jingkang50 Merge pull request #44 from Jingkang50/dev_jkyang ... aea7e8d 10 days ago 100 commits

| | | |
|---|---|---|
| assets | update readme | 3 months ago |
| configs | update fsood | 11 days ago |
| openood | update fsood | 11 days ago |
| scripts | update fsood | 11 days ago |
| tools | fix covid | 17 days ago |
| .gitignore | load fsood | 12 days ago |
| .pre-commit-config.yaml | fix mds | 3 months ago |
| LICENSE | Initial commit | 5 months ago |
| README.md | update readme | 10 days ago |
| codespell_ignored.txt | rename codespell | 3 months ago |
| environment.yml | update fsood | 11 days ago |
| main.py | fix fsood | 3 months ago |

README.md

## OpenOOD: Benchmarking Generalized OOD Detection

This repository reproduces representative methods within the `Generalized Out-of-Distribution Detection Framework`, aiming to make a fair comparison across methods that initially developed for anomaly detection, novelty detection, open set recognition, and out-of-distribution detection. This codebase is still under construction. Comments, issues, contributions, and collaborations are all welcomed!

**About**

Benchmarking Generalized Out-of-Distribution Detection

Readme
MIT License
39 stars
4 watching
3 forks

**Releases**

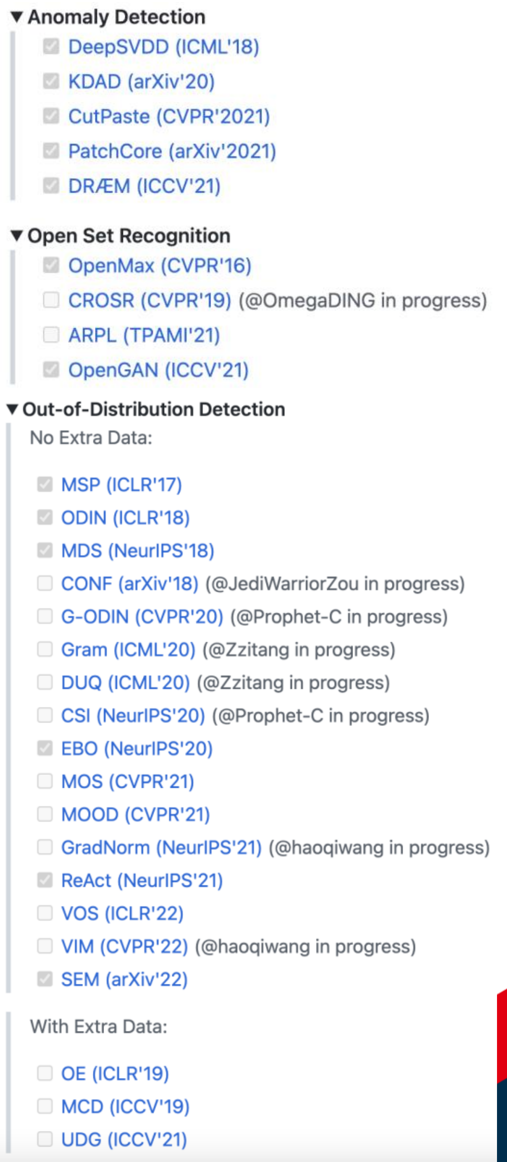
No releases published
Create a new release

**Packages**

No packages published
Publish your first package

**Contributors** 3

Jingkang50 Jingkang Yang
Prophet-C Pengyun Wang
JediWarriorZou DEJIAN ZOU



▼ **Anomaly Detection**

- ☑ DeepSVDD (ICML'18)
- ☑ KDAD (arXiv'20)
- ☑ CutPaste (CVPR'2021)
- ☑ PatchCore (arXiv'2021)
- ☑ DRÆM (ICCV'21)

▼ **Open Set Recognition**

- ☑ OpenMax (CVPR'16)
- ☐ CROSR (CVPR'19) (@OmegaDING in progress)
- ☐ ARPL (TPAMI'21)
- ☑ OpenGAN (ICCV'21)

▼ **Out-of-Distribution Detection**

No Extra Data:

- ☑ MSP (ICLR'17)
- ☑ ODIN (ICLR'18)
- ☑ MDS (NeurIPS'18)
- ☐ CONF (arXiv'18) (@JediWarriorZou in progress)
- ☐ G-ODIN (CVPR'20) (@Prophet-C in progress)
- ☐ Gram (ICML'20) (@Zzitang in progress)
- ☐ DUQ (ICML'20) (@Zzitang in progress)
- ☐ CSI (NeurIPS'20) (@Prophet-C in progress)
- ☑ EBO (NeurIPS'20)
- ☐ MOS (CVPR'21)
- ☐ MOOD (CVPR'21)
- ☐ GradNorm (NeurIPS'21) (@haoqiwang in progress)
- ☑ ReAct (NeurIPS'21)
- ☐ VOS (ICLR'22)
- ☐ VIM (CVPR'22) (@haoqiwang in progress)
- ☑ SEM (arXiv'22)

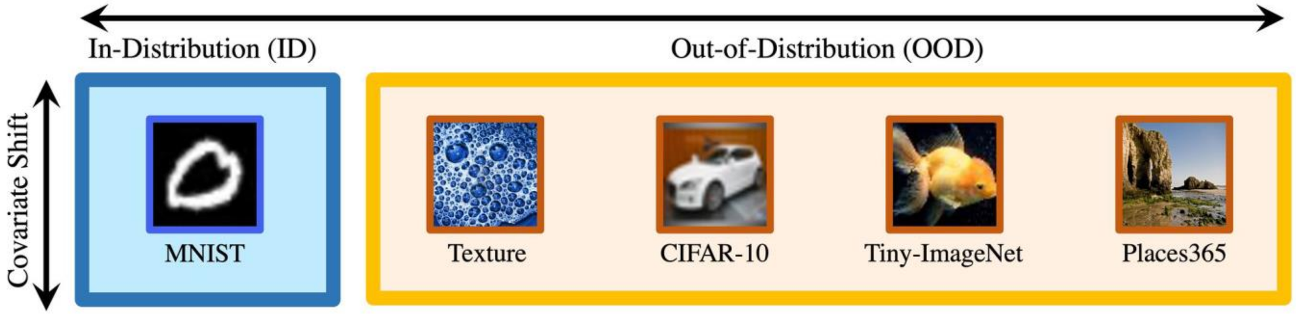
With Extra Data:

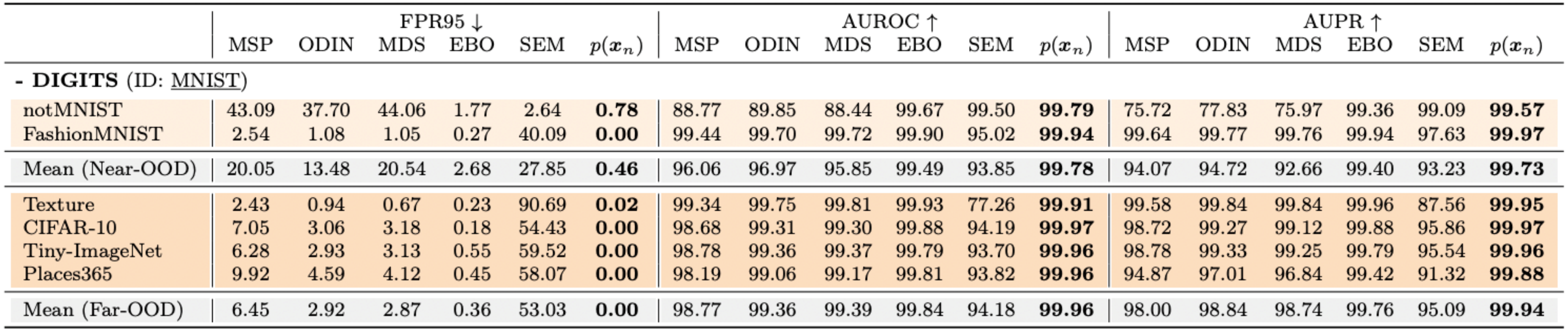
- ☐ OE (ICLR'19)
- ☐ MCD (ICCV'19)
- ☐ UDG (ICCV'21)

# Problem with Classic OOD Benchmark

## Problem on current OOD Benchmarks

- **Classic OOD Benchmark:**
  - Saturated benchmark
  - Model can only rely on covariate shift detection to performing OOD detection
  - But OOD detection should focus on semantic anomalies





| | FPR95 ↓ | | | | | | AUROC ↑ | | | | | | AUPR ↑ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSP | ODIN | MDS | EBO | SEM | $p(\boldsymbol{x}_n)$ | MSP | ODIN | MDS | EBO | SEM | $p(\boldsymbol{x}_n)$ | MSP | ODIN | MDS | EBO | SEM | $p(\boldsymbol{x}_n)$ |
| - **DIGITS** (ID: MNIST) | | | | | | | | | | | | | | | | | | |
| notMNIST | 43.09 | 37.70 | 44.06 | 1.77 | 2.64 | **0.78** | 88.77 | 89.85 | 88.44 | 99.67 | 99.50 | **99.79** | 75.72 | 77.83 | 75.97 | 99.36 | 99.09 | **99.57** |
| FashionMNIST | 2.54 | 1.08 | 1.05 | 0.27 | 40.09 | **0.00** | 99.44 | 99.70 | 99.72 | 99.90 | 95.02 | **99.94** | 99.64 | 99.77 | 99.76 | 99.94 | 97.63 | **99.97** |
| Mean (Near-OOD) | 20.05 | 13.48 | 20.54 | 2.68 | 27.85 | **0.46** | 96.06 | 96.97 | 95.85 | 99.49 | 93.85 | **99.78** | 94.07 | 94.72 | 92.66 | 99.40 | 93.23 | **99.73** |
| Texture | 2.43 | 0.94 | 0.67 | 0.23 | 90.69 | **0.02** | 99.34 | 99.75 | 99.81 | 99.93 | 77.26 | **99.91** | 99.58 | 99.84 | 99.84 | 99.96 | 87.56 | **99.95** |
| CIFAR-10 | 7.05 | 3.06 | 3.18 | 0.18 | 54.43 | **0.00** | 98.68 | 99.31 | 99.30 | 99.88 | 94.19 | **99.97** | 98.72 | 99.27 | 99.12 | 99.88 | 95.86 | **99.97** |
| Tiny-ImageNet | 6.28 | 2.93 | 3.13 | 0.55 | 59.52 | **0.00** | 98.78 | 99.36 | 99.37 | 99.79 | 93.70 | **99.96** | 98.78 | 99.33 | 99.25 | 99.79 | 95.54 | **99.96** |
| Places365 | 9.92 | 4.59 | 4.12 | 0.45 | 58.07 | **0.00** | 98.19 | 99.06 | 99.17 | 99.81 | 93.82 | **99.96** | 94.87 | 97.01 | 96.84 | 99.42 | 91.32 | **99.88** |
| Mean (Far-OOD) | 6.45 | 2.92 | 2.87 | 0.36 | 53.03 | **0.00** | 98.77 | 99.36 | 99.39 | 99.84 | 94.18 | **99.96** | 98.00 | 98.84 | 98.74 | 99.76 | 95.09 | **99.94** |

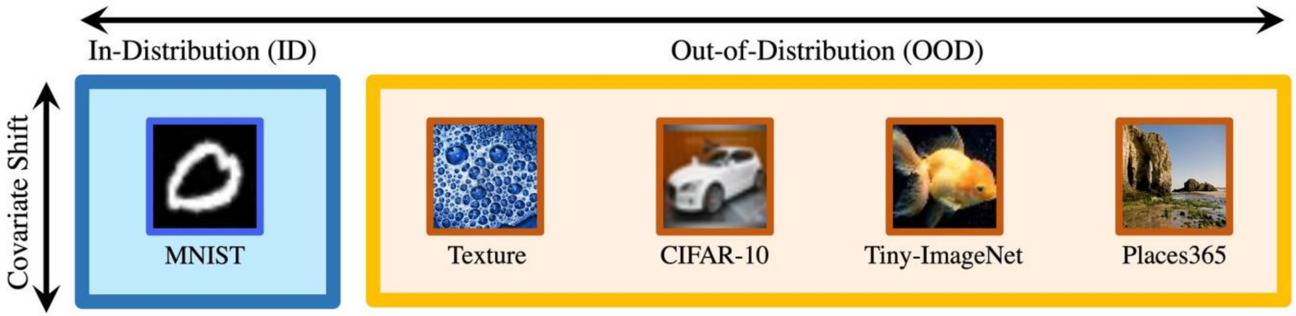
Table: Results on Standard OOD Detection Benchmarks

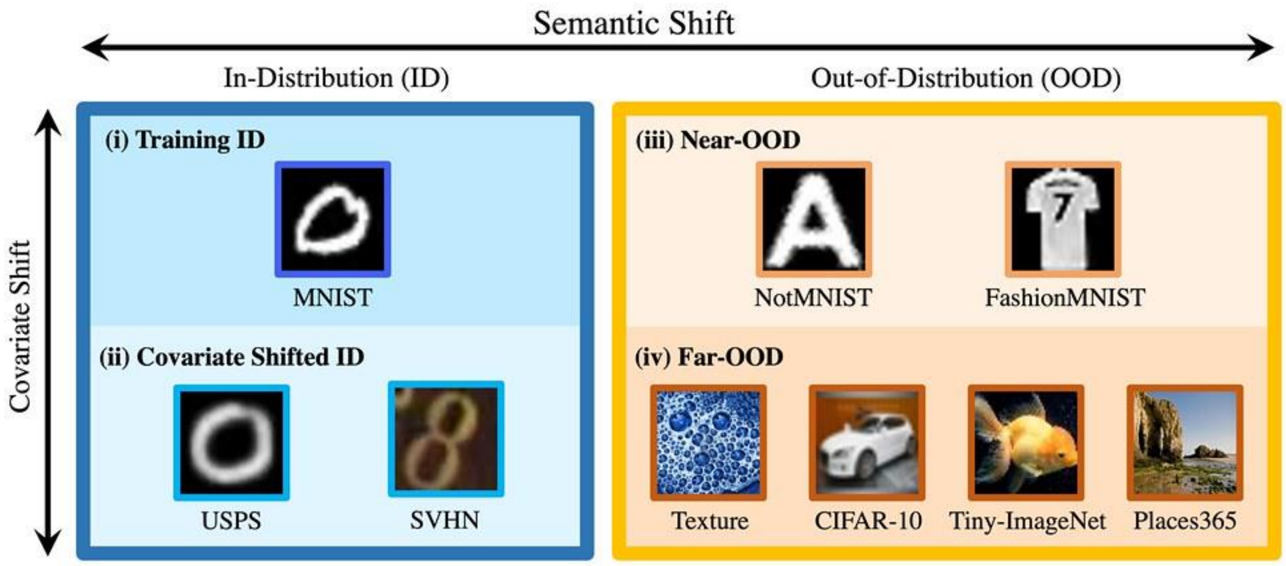
# Full-Spectrum OOD Benchmark

## Problem on current OOD Benchmarks

- **Classic OOD Benchmark:**
    - Saturated benchmark
    - Model can only rely on covariate shift detection to performing OOD detection
    - But OOD detection should focus on semantic anomalies

- **Full-Spectrum OOD Benchmark:**
    - Introducing Covariate-Shifted In-Distribution Data
    - A better benchmark to evaluate semantic shift detection capability
    - Promoting robustness in OOD detection





# Full-Spectrum OOD Benchmark

**Full-Spectrum OOD Benchmark:**

- Introducing Covariate-Shifted In-Distribution Data
- A better benchmark to evaluate semantic shift detection capability
- Promoting robustness in OOD detection

- Most previous methods completely fail on FS-OOD setting
- In fact, CIFAR-level OOD detection benchmarks are still not saturated and may still need more exploration

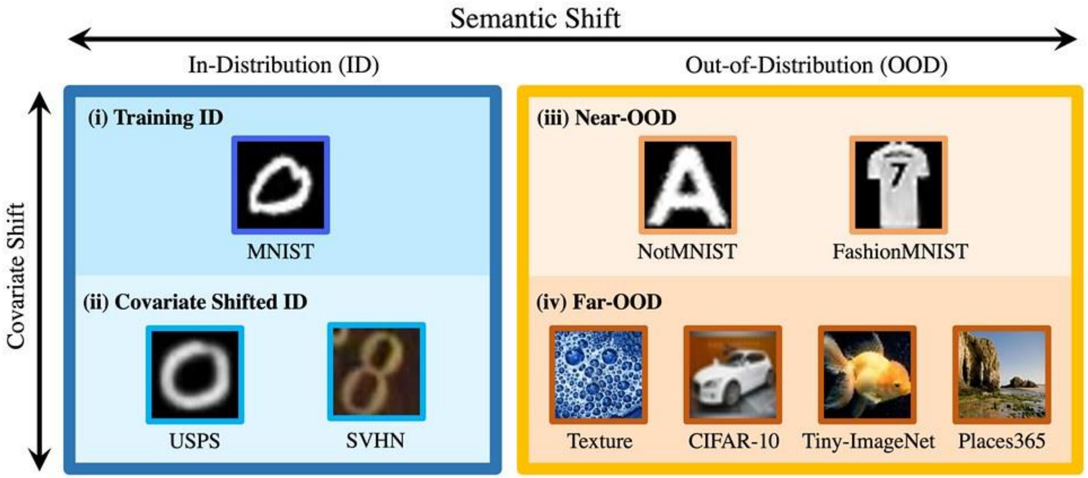
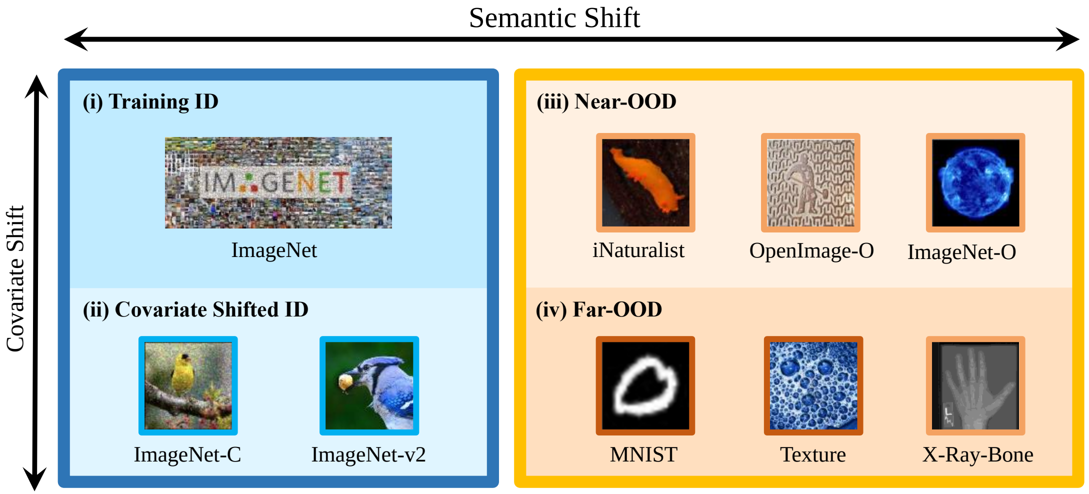


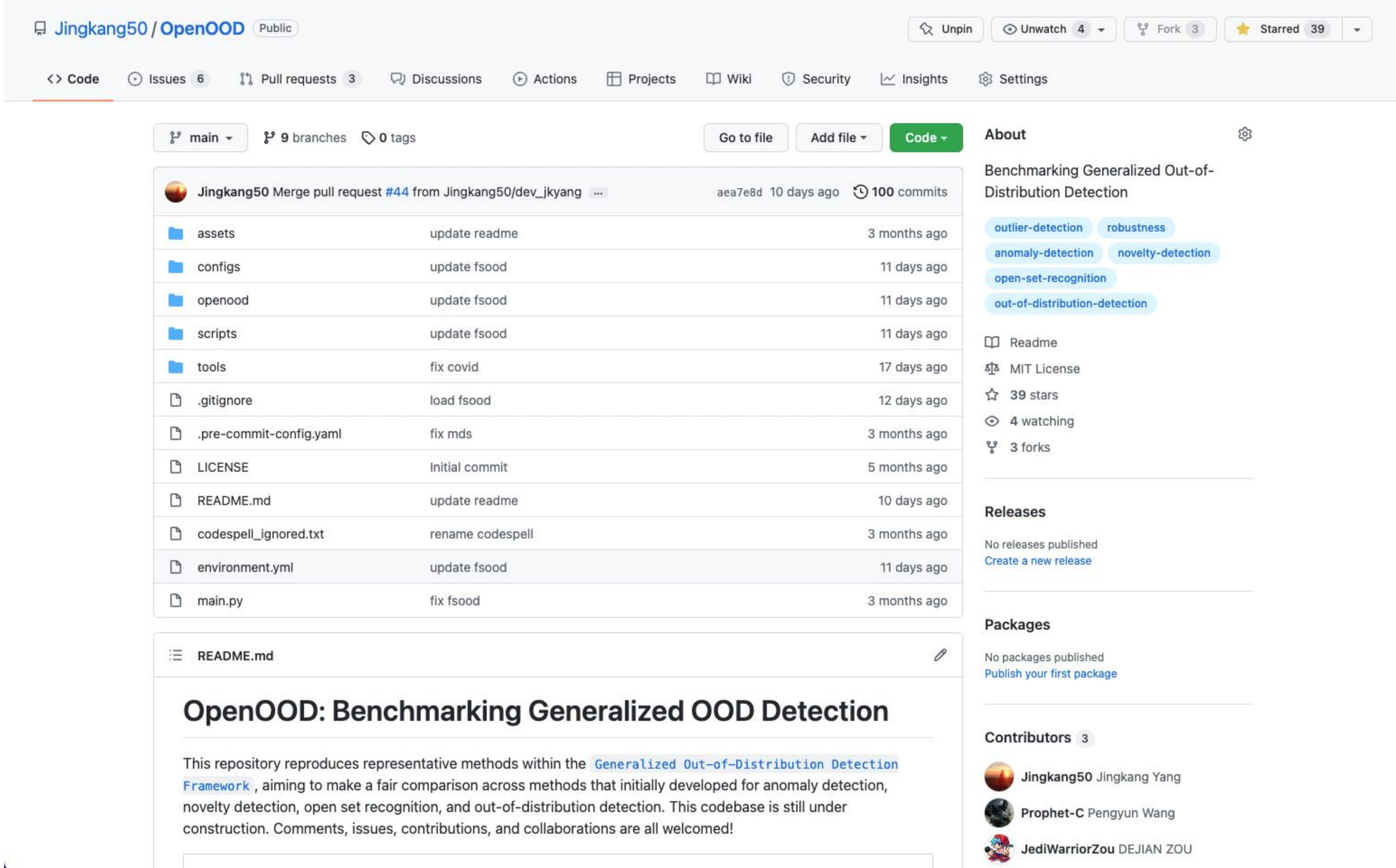
**Figure**: Large-Scale Full-Spectrum OOD Detection Benchmarks



[Jingkang Yang, Kaiyang Zhou, Ziwei Liu. Full-Spectrum OOD Detection. arXiv:2114.05306. 2022 ]

# Code, Models & Dataset

**OpenOOD:** https://github.com/Jingkang50/OpenOOD



Jingkang50 Merge pull request #44 from Jingkang50/dev_jkyang … aea7e8d 10 days ago 100 commits

| | | |
|---|---|---|
| assets | update readme | 3 months ago |
| configs | update fsood | 11 days ago |
| openood | update fsood | 11 days ago |
| scripts | update fsood | 11 days ago |
| tools | fix covid | 17 days ago |
| .gitignore | load fsood | 12 days ago |
| .pre-commit-config.yaml | fix mds | 3 months ago |
| LICENSE | Initial commit | 5 months ago |
| README.md | update readme | 10 days ago |
| codespell_ignored.txt | rename codespell | 3 months ago |
| environment.yml | update fsood | 11 days ago |
| main.py | fix fsood | 3 months ago |

README.md

## OpenOOD: Benchmarking Generalized OOD Detection

This repository reproduces representative methods within the `Generalized Out-of-Distribution Detection Framework`, aiming to make a fair comparison across methods that initially developed for anomaly detection, novelty detection, open set recognition, and out-of-distribution detection. This codebase is still under construction. Comments, issues, contributions, and collaborations are all welcomed!

**About**

Benchmarking Generalized Out-of-Distribution Detection

outlier-detection robustness anomaly-detection novelty-detection open-set-recognition out-of-distribution-detection

- Readme
- MIT License
- 39 stars
- 4 watching
- 3 forks

**Releases**

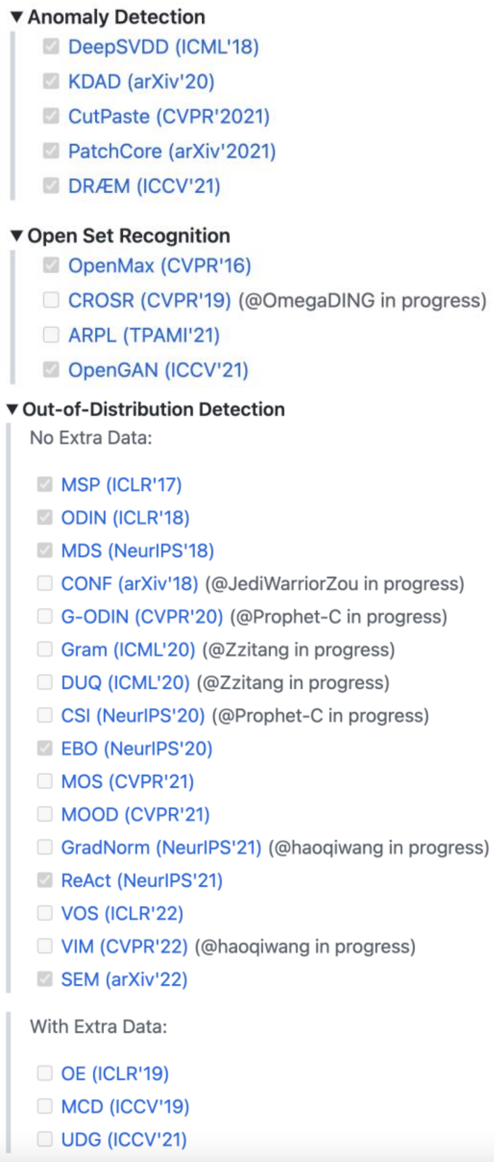
No releases published
Create a new release

**Packages**

No packages published
Publish your first package

**Contributors** 3

- **Jingkang50** Jingkang Yang
- **Prophet-C** Pengyun Wang
- **JediWarriorZou** DEJIAN ZOU



▼ **Anomaly Detection**
- ☑ DeepSVDD (ICML'18)
- ☑ KDAD (arXiv'20)
- ☑ CutPaste (CVPR'2021)
- ☑ PatchCore (arXiv'2021)
- ☑ DRÆM (ICCV'21)

▼ **Open Set Recognition**
- ☑ OpenMax (CVPR'16)
- ☐ CROSR (CVPR'19) (@OmegaDING in progress)
- ☐ ARPL (TPAMI'21)
- ☑ OpenGAN (ICCV'21)

▼ **Out-of-Distribution Detection**

No Extra Data:

- ☑ MSP (ICLR'17)
- ☑ ODIN (ICLR'18)
- ☑ MDS (NeurIPS'18)
- ☐ CONF (arXiv'18) (@JediWarriorZou in progress)
- ☐ G-ODIN (CVPR'20) (@Prophet-C in progress)
- ☐ Gram (ICML'20) (@Zzitang in progress)
- ☐ DUQ (ICML'20) (@Zzitang in progress)
- ☐ CSI (NeurIPS'20) (@Prophet-C in progress)
- ☑ EBO (NeurIPS'20)
- ☐ MOS (CVPR'21)
- ☐ MOOD (CVPR'21)
- ☐ GradNorm (NeurIPS'21) (@haoqiwang in progress)
- ☑ ReAct (NeurIPS'21)
- ☐ VOS (ICLR'22)
- ☐ VIM (CVPR'22) (@haoqiwang in progress)
- ☑ SEM (arXiv'22)

With Extra Data:

- ☐ OE (ICLR'19)
- ☐ MCD (ICCV'19)
- ☐ UDG (ICCV'21)

# Thank you for listening!

**Semantic Shift**

***OOD Detection***

***Zero-shot / Few-shot / Long-tailed Learning***

**Sensory Modalities**

**Neural Architectures**

***Corruptions / Perturbations / Domain Shifts***

**Covariate Shift**