# Few-Shot Object Detection via Association and DIscrimination

Yuhang Cao[1] Jiaqi Wang[1]✉ Ying Jin[1] Tong Wu[1]
Kai Chen[3,4] Ziwei Liu[2] Dahua Lin[1,4]
[1]The Chinese University of Hong Kong
[2]S-Lab, Nanyang Technological University [3]SenseTime Research
[4]Shanghai AI Laboratory

# Few-Shot Object Detection (FSOD)

Abundant base dataset of base classes $C^B$

Scarce novel dataset of novel classes $C^N$

Train

Few Shot Detector

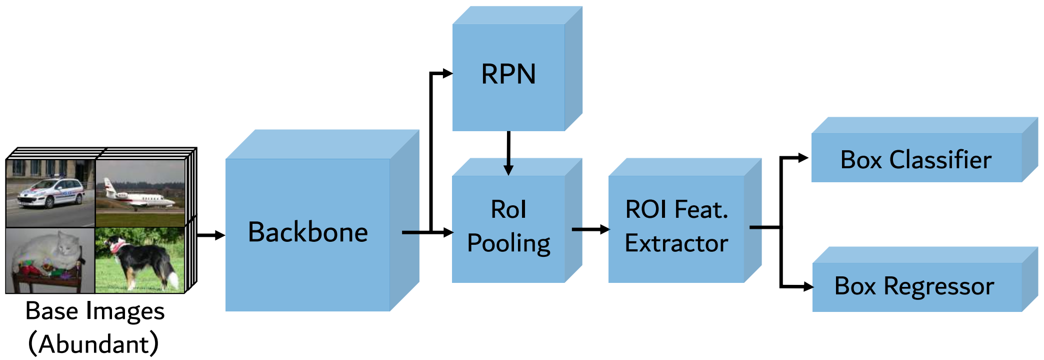
Test

Detect objects of both $C^B \cup C^N$ in the test dataset

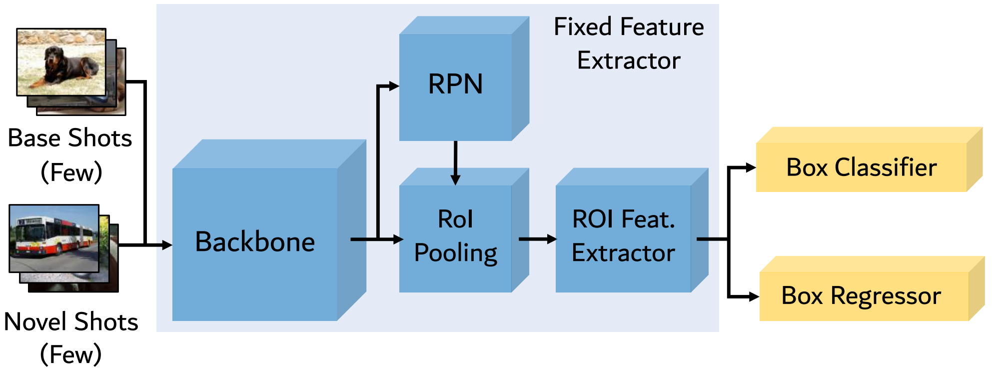
*N*-way-*K*-shot: *N* novel classes, each novel class has *K* annotated instances

# Fine-turning-based Few-Shot Detector

Stage I: Base training




Stage II: Few-shot fine-tuning




Base set $D^B$ of classes $C^B$:

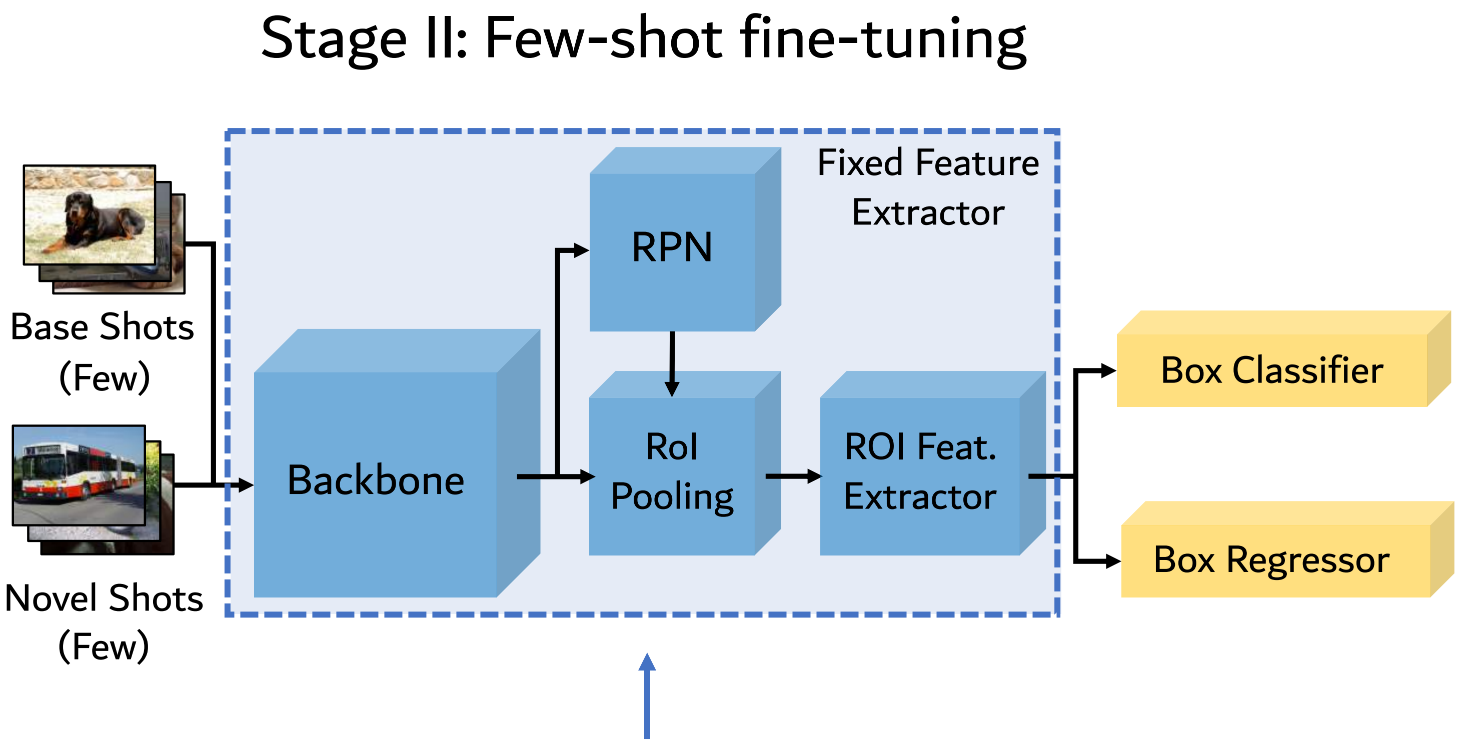
- Training feature extrator
- Training box predictor

Balanced set of classes $C^B \cup C^N$:

- Fix feature extrator
- Fine-tune box predictor

Frustratingly Simple Few-Shot Object Detection (TFA) (ICML 2020)

# Philosafy of the design of ft-based pipeline




- Class-agnostic components
- Encode rich base knowledge

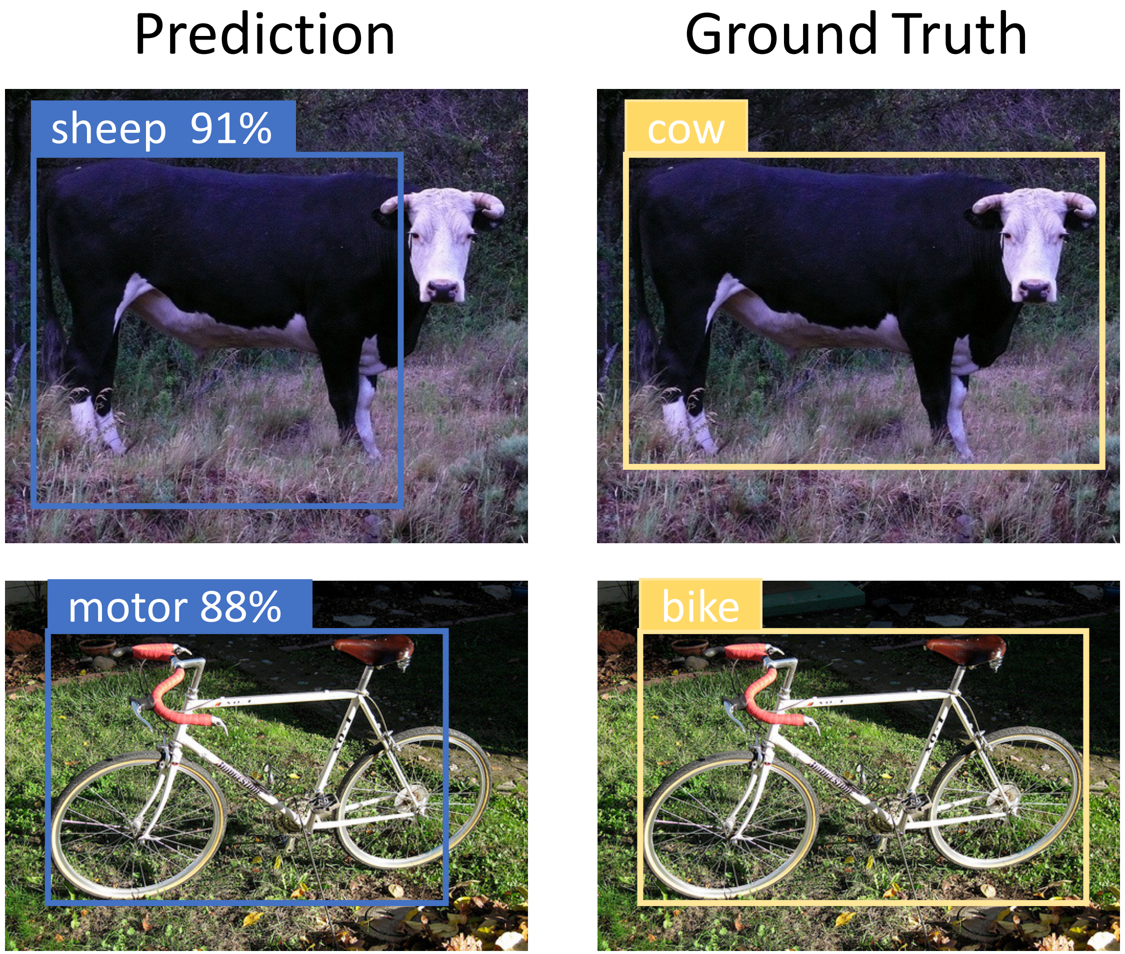
- Avoid over-fitting on small novel set $D^N$

Frustratingly Simple Few-Shot Object Detection (ICML 2020)

# Evil: Misclassification

- Fixed feature extractor can yield similar feature representation of texture similar objects

- The box classifier (a single fc) is not able to accurately classify similar objects



# Motivation

1. Novel class cow is similar to **single** base class sheep

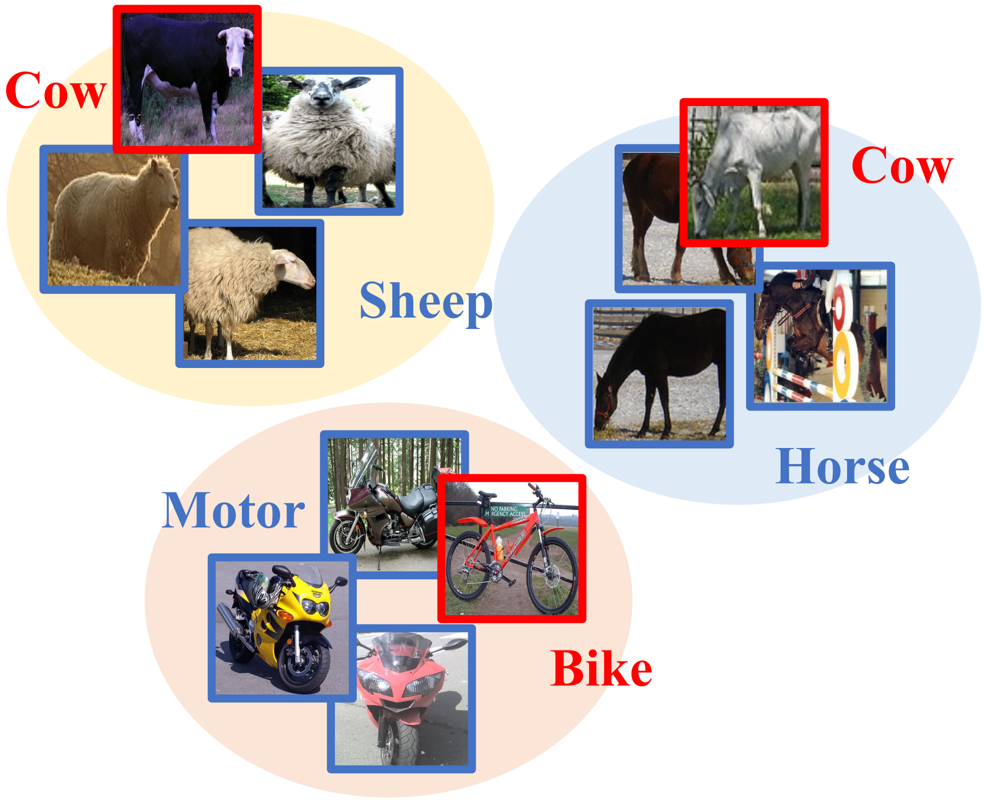
→ Feature space of cow overlaps with sheep

→ <u>Small inter-class separability</u>

2. Novel class cow is similar to **two** base classes sheep and horse

→ Feature space of cow scatters across sheep and horse

→ <u>Large intra-class variances</u>




Ellipse: feature space of a base class

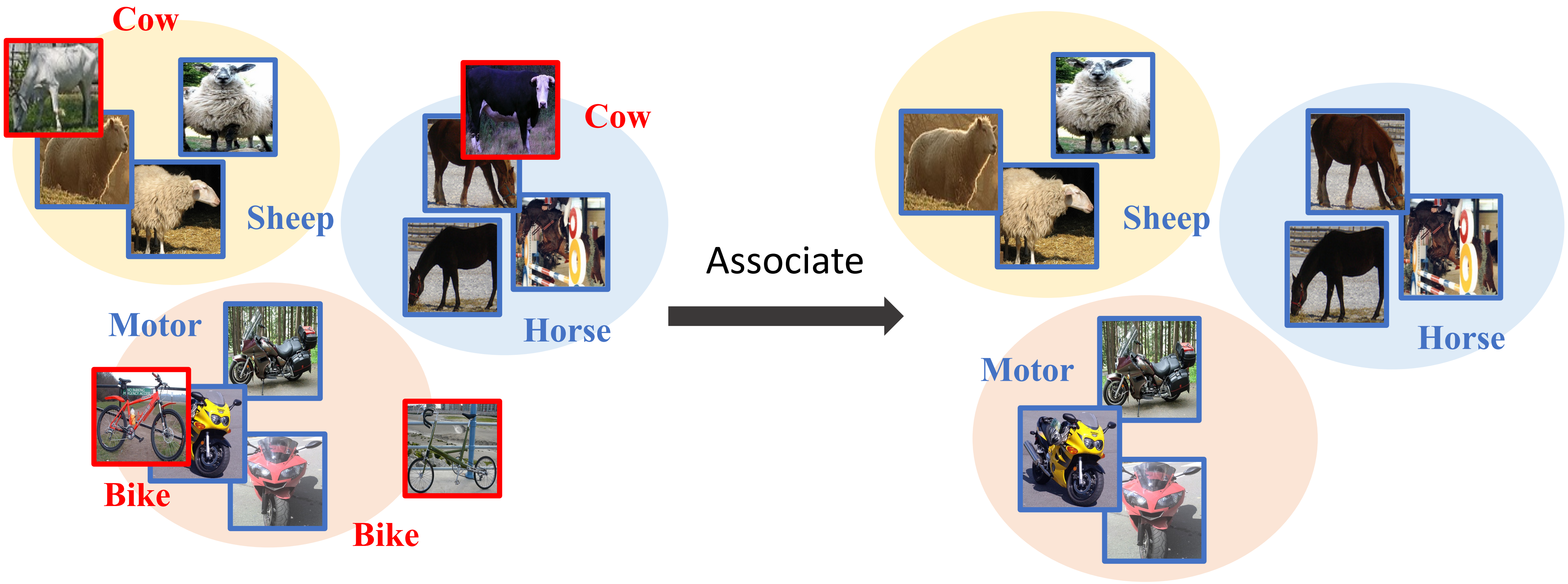
# Our method: FADI

To alleviate the limitations, we propose a two-step fine-tuning framework:

1. ***Association***: **compact intra-class structure**
   - Similarity Measurement
   - Feature Distribution Alignment

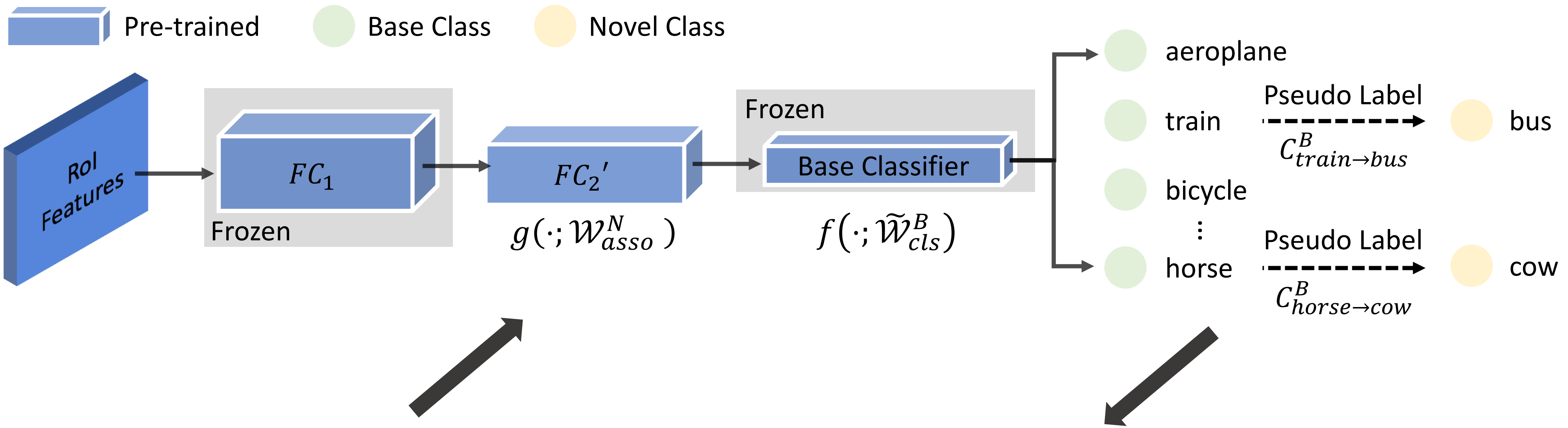
2. ***Discrimination***: **ensure enough inter-class separability**
   - Disentangling
   - Set-Specialized Margin Loss

# Conceptualization of Association




Align the feature distribution of each novel class with its most semantically similar class

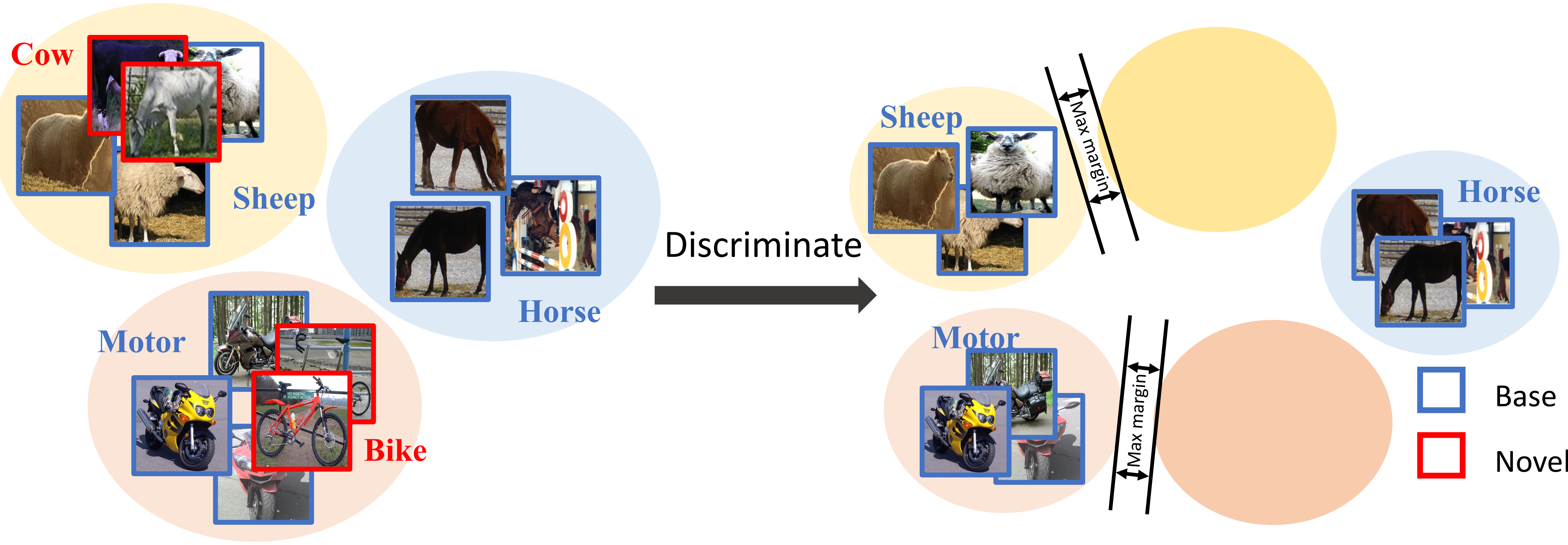
# Instantiation of Association




3. The classifier identifies novel class as base class
4. Features of novel class shift toward its associated base class

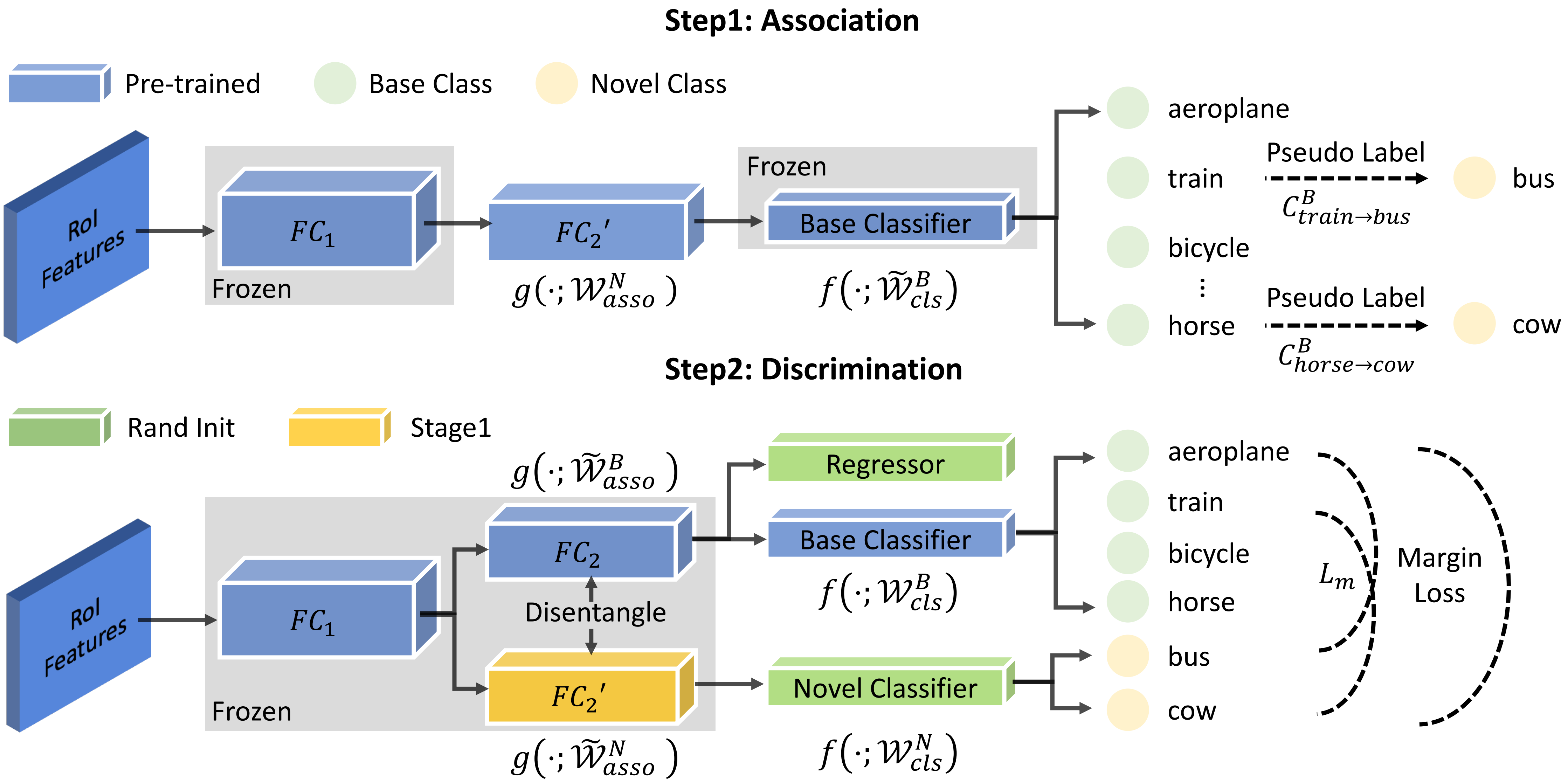
1. Associate each novel class with its most sematically similar base class
2. Replace the label of novel class with its associated base label

# Conceptualization of Discrimination




Separate the associated base and novel classes by disentangling and margin loss

# Instantiation of Discrimination



# Set-Specialized Margin Loss

Cosine classifier: adopt cosine similarity to formulate the logit prediction

$$p_{y_i} = \frac{\tau \cdot \mathbf{x}^T \mathcal{W}_{y_i}}{||\mathbf{x}|| \cdot ||\mathcal{W}_{y_i}||}, \quad s_{y_i} = \frac{e^{p_{y_i}}}{\sum_{j=1}^{C} e^{p_j}},$$

Maximizing the score difference of different classes

$$\mathcal{L}_{m_i} = \sum_{j=1, j \neq y_i}^{C} -\log((\underline{s_{y_i} - s_j})^+ + \epsilon),$$

Inter-class margin: $s_{y_i} - s_j$

# Set-Specialized Margin Loss

Maximizing the score difference of different classes

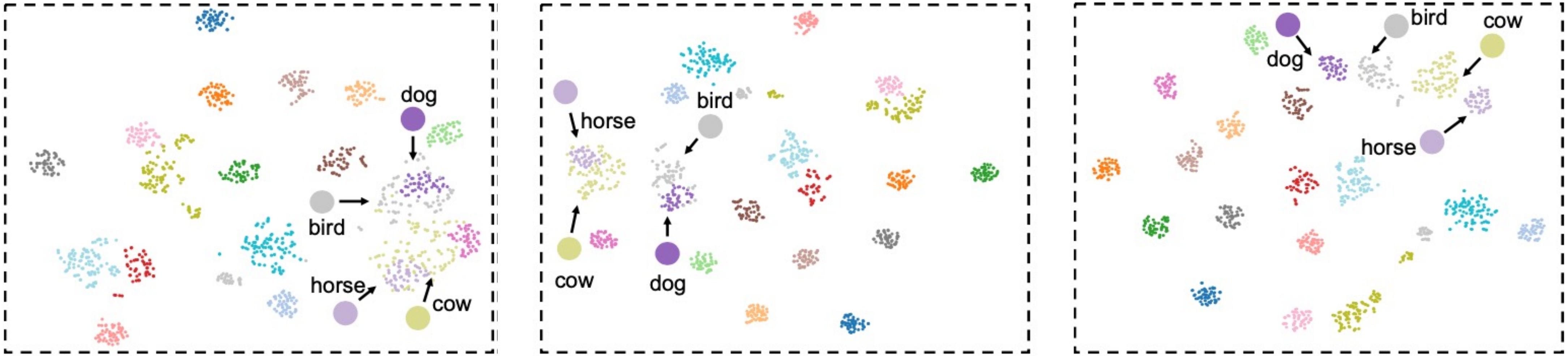
$$\mathcal{L}_{m_i} = \sum_{j=1, j \neq y_i}^{C} -\log((s_{y_i} - s_j)^+ + \epsilon),$$

Introducing different margin to different class set

$$\mathcal{L}_m = \boxed{\sum_{\{i|y_i \in C^B\}} \alpha \cdot \mathcal{L}_{m_i}} + \boxed{\sum_{\{i|y_i \in C^N\}} \beta \cdot \mathcal{L}_{m_i}} + \boxed{\sum_{\{i|y_i = C^0\}} \gamma \cdot \mathcal{L}_{m_i}},$$

$C^B$: base classes; $C^N$: novel classes $C^0$: background classes

# Effectiveness of FADI




(a) TFA (b) Association (c) Discrimination

t-SNE visualization of feature distribution of TFA and our FADI

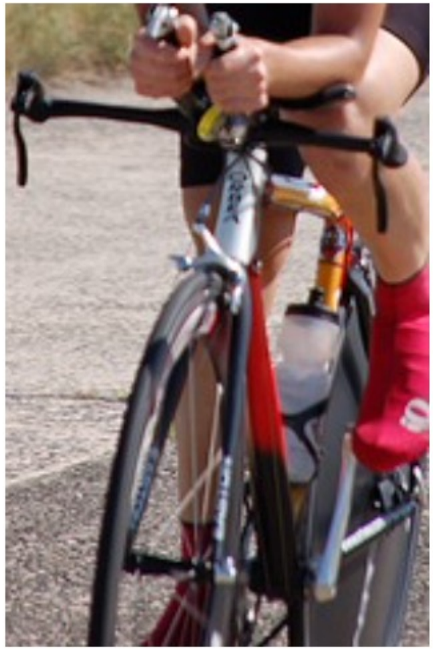
# Overall Performance on Pascal VOC

| Method / Shot | Backbone | Novel Split 1 | | | | | Novel Split 2 | | | | | Novel Split 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 |
| LSTD [2] | VGG-16 | 8.2 | 1.0 | 12.4 | 29.1 | 38.5 | 11.4 | 3.8 | 5.0 | 15.7 | 31.0 | 12.6 | 8.5 | 15.0 | 27.3 | 36.3 |
| YOLOv2-ft [29] | YOLO V2 | 6.6 | 10.7 | 12.5 | 24.8 | 38.6 | 12.5 | 4.2 | 11.6 | 16.1 | 33.9 | 13.0 | 15.9 | 15.0 | 32.2 | 38.4 |
| †FSRW [12] | YOLO V2 | 14.8 | 15.5 | 26.7 | 33.9 | 47.2 | 15.7 | 15.3 | 22.7 | 30.1 | 40.5 | 21.3 | 25.6 | 28.4 | 42.8 | 45.9 |
| †MetaDet [29] | YOLO V2 | 17.1 | 19.1 | 28.9 | 35.0 | 48.8 | 18.2 | 20.6 | 25.9 | 30.6 | 41.5 | 20.1 | 22.3 | 27.9 | 41.9 | 42.9 |
| †RepMet [13] | InceptionV3 | 26.1 | 32.9 | 34.4 | 38.6 | 41.3 | 17.2 | 22.1 | 23.4 | 28.3 | 35.8 | 27.5 | 31.1 | 31.5 | 34.4 | 37.2 |
| FRCN-ft [29] | FRCN-R101 | 13.8 | 19.6 | 32.8 | 41.5 | 45.6 | 7.9 | 15.3 | 26.2 | 31.6 | 39.1 | 9.8 | 11.3 | 19.1 | 35.0 | 45.1 |
| FRCN+FPN-ft [27] | FRCN-R101 | 8.2 | 20.3 | 29.0 | 40.1 | 45.5 | 13.4 | 20.6 | 28.6 | 32.4 | 38.8 | 19.6 | 20.8 | 28.7 | 42.2 | 42.1 |
| †MetaDet [29] | FRCN-R101 | 18.9 | 20.6 | 30.2 | 36.8 | 49.6 | 21.8 | 23.1 | 27.8 | 31.7 | 43.0 | 20.6 | 23.9 | 29.4 | 43.9 | 44.1 |
| †Meta R-CNN [32] | FRCN-R101 | 19.9 | 25.5 | 35.0 | 45.7 | 51.5 | 10.4 | 19.4 | 29.6 | 34.8 | 45.4 | 14.3 | 18.2 | 27.5 | 41.2 | 48.1 |
| TFA w/ fc [27] | FRCN-R101 | 36.8 | 29.1 | 43.6 | 55.7 | 57.0 | 18.2 | 29.0 | 33.4 | 35.5 | 39.0 | 27.7 | 33.6 | 42.5 | 48.7 | 50.2 |
| TFA w/ cos [27] | FRCN-R101 | 39.8 | 36.1 | 44.7 | 55.7 | 56.0 | 23.5 | 26.9 | 34.1 | 35.1 | 39.1 | 30.8 | 34.8 | 42.8 | 49.5 | 49.8 |
| MPSR [30] | FRCN-R101 | 41.7 | - | 51.4 | 55.2 | 61.8 | 24.4 | - | 39.2 | 39.9 | 47.8 | 35.6 | - | 42.3 | 48.0 | 49.7 |
| SRR-FSD [33] | FRCN-R101 | 47.8 | 50.5 | 51.3 | 55.2 | 56.8 | **32.5** | **35.3** | 39.1 | 40.8 | 43.8 | 40.1 | 41.5 | 44.3 | 46.9 | 46.4 |
| FSCE [22] | FRCN-R101 | 44.2 | 43.8 | 51.4 | **61.9** | **63.4** | 27.3 | 29.5 | **43.5** | **44.2** | **50.2** | 37.2 | 41.9 | 47.5 | 54.6 | 58.5 |
| FADI (Ours) | FRCN-R101 | **50.3** | **54.8** | **54.2** | 59.3 | 63.2 | 30.6 | 35.0 | 40.3 | 42.8 | 48.0 | **45.7** | **49.7** | **49.1** | **55.0** | **59.6** |

New SOTA on shot 1, 2, 3 and 1, 2, 3, 5, 10 on split1 and 3, respectively

# Superiority of semantic similarity over visual similarity

A cat sits on a chair



A human rides a bicycle

- Co-occurrence can yield misleading visual similarity.
- Text semantic similarity is regardless of co-occurrance.

# Thank you!