



# Incrementally Zero-Shot Detection by an Extreme Value Analyzer

Sixiao Zheng, Yanwei Fu, Yanxi Hou
Fudan University

# Motivation

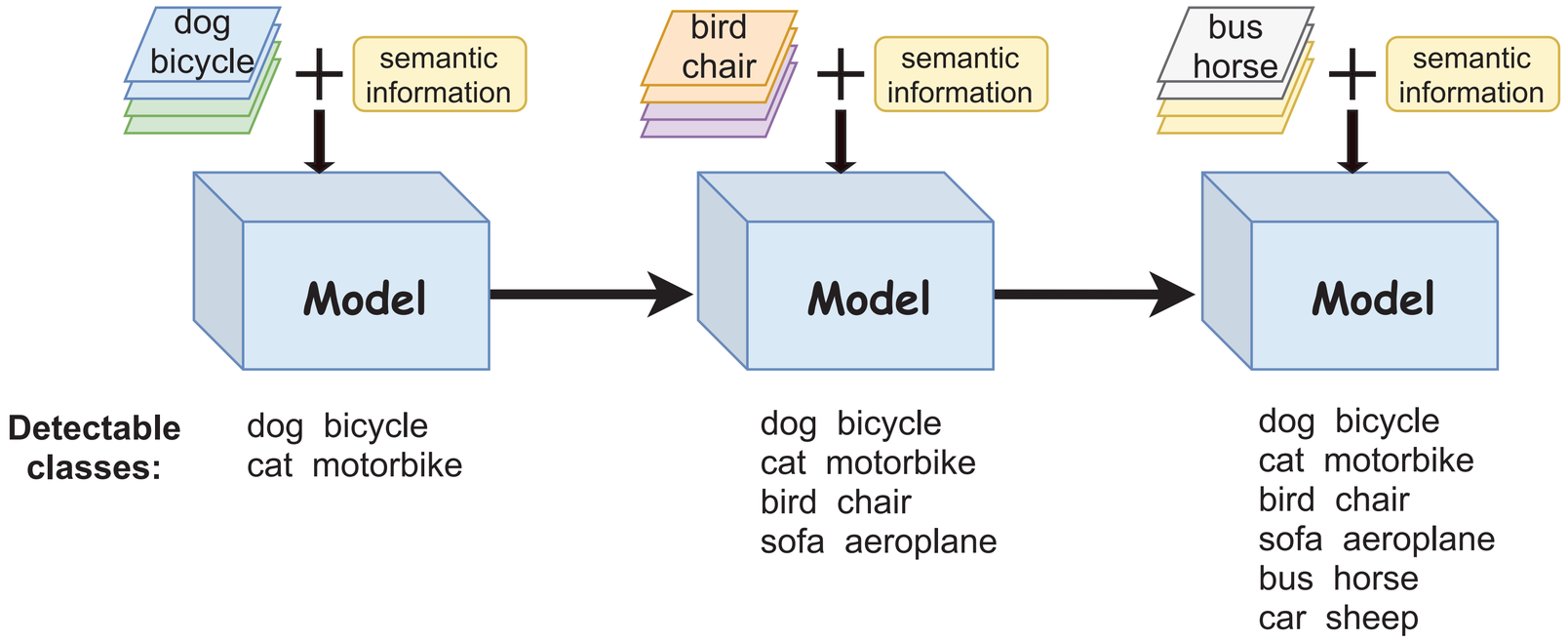
- Most successful object detection models are formulated as supervised learning problems in a batch setting.

- Humans not only have the ability to recognize novel unseen object classes, but incrementally incorporate these novel object classes to existing knowledge preserved as well.

# Contributions

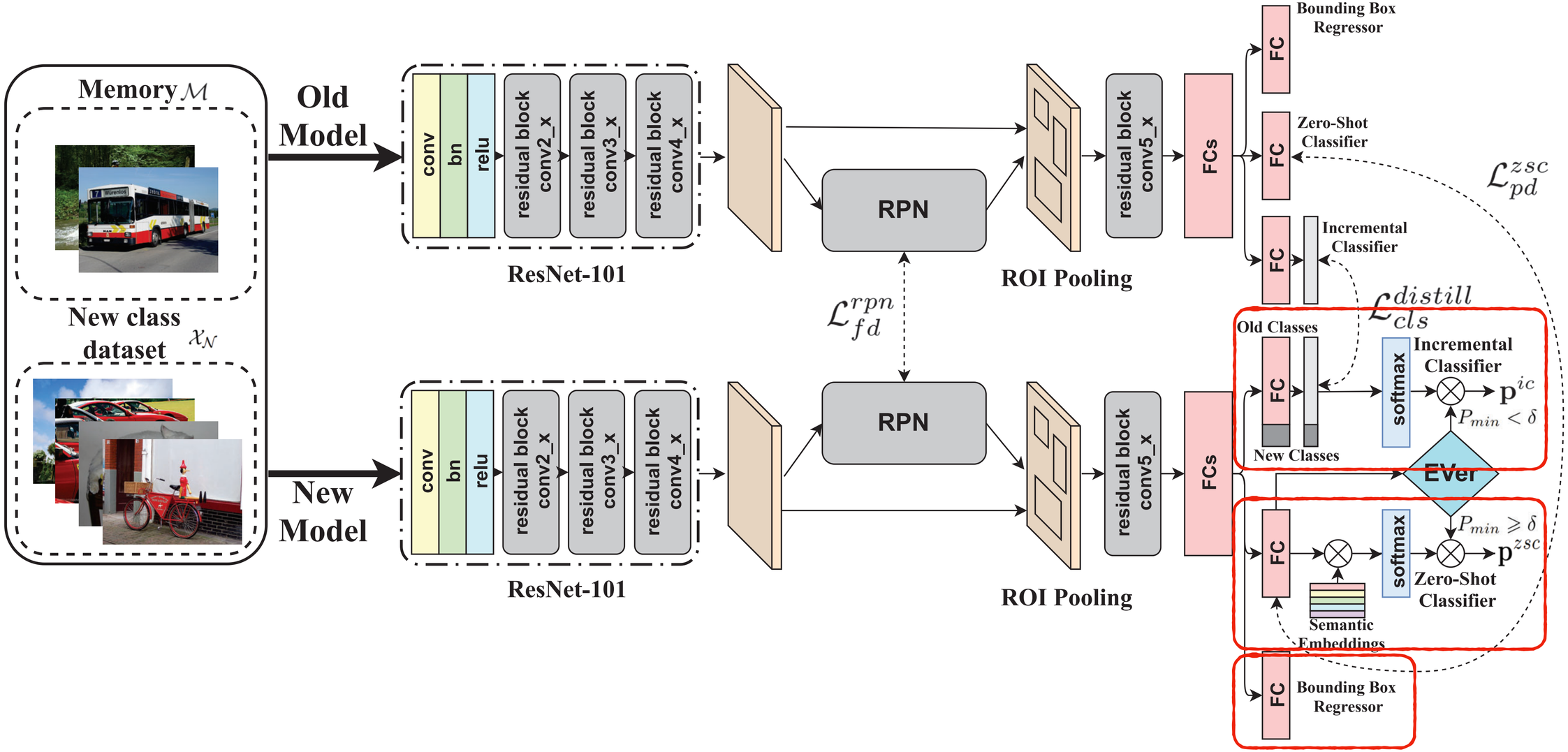
- We introduce a new task of Incrementally Zero-Shot Detection (IZSD) for real-world object detection.
- We propose an innovative model – IZSD-EVer in addressing IZSD by integrating Zero-shot detection (ZSD) and Class-Incremental Detection (CID) in a single model.
- We propose Extreme Value Analyzer (EVer), which is the first time Extreme Value Theory (EVT) has been introduced into object detection.
- We propose two novel losses, i.e., backgroud-forground MSE loss (bfMSE) and projection distance (PD) loss.

# New task: IZSD

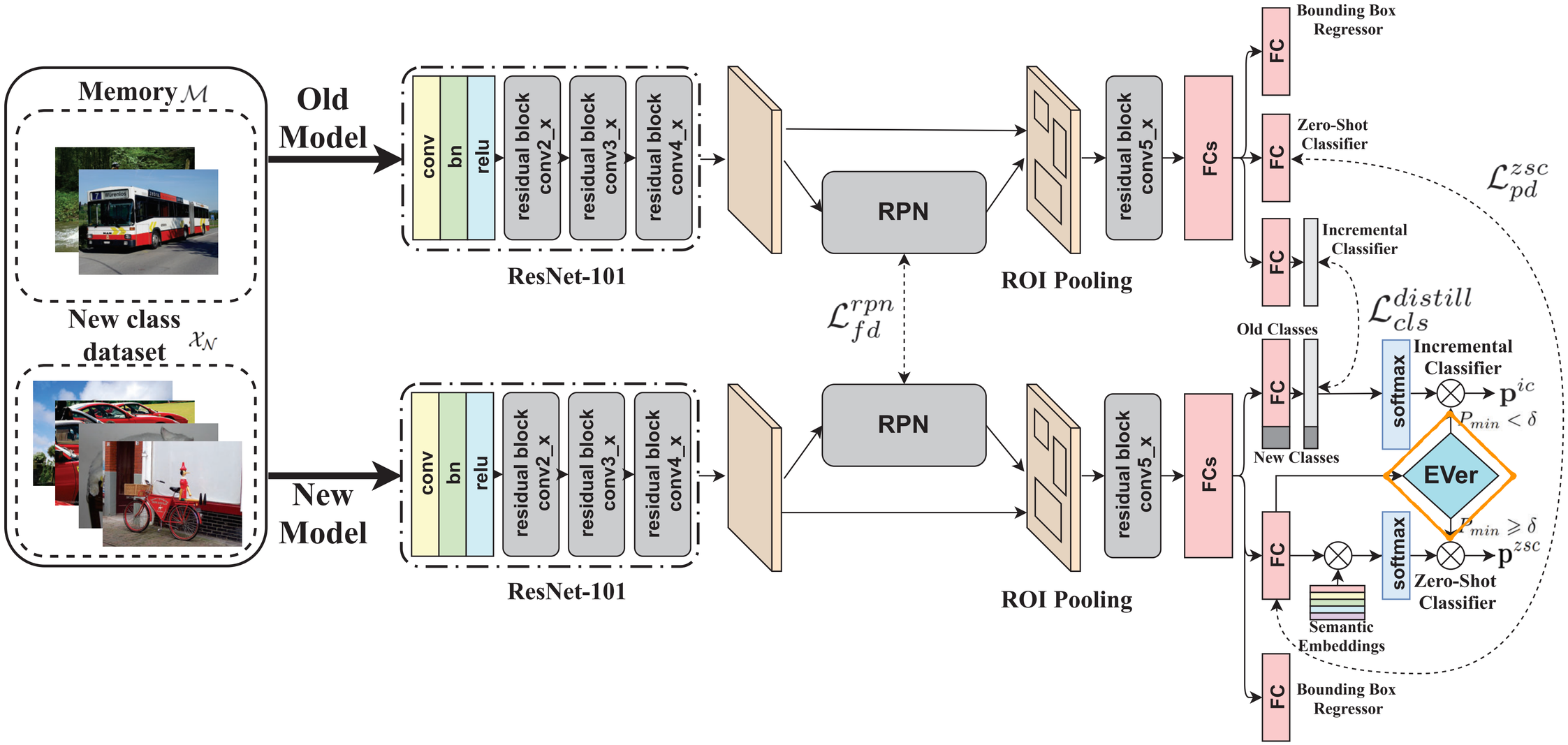



- examples of new seen object classes and semantic information are utilized to update the detection model
- **Goal**: to detect objects from old seen, new seen, and unseen classes.
- **Challenges**: efficient in semantic knowledge transfer, robust to catastrophic forgetting

# Proposed Model: IZSD-EVer



# Proposed Model: IZSD-EVer



# Backbone with Semantic Embedding

- **bfMSE Loss**: $$\mathcal{L}_{bfmse} = \frac{1}{N_{bg}} \sum_{i=1}^{N_{bg}} \mathbb{I}_i^{bg} \|\mathbf{s}_i - \mathbf{E}_{0,:}\|_2^2 + \alpha \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} \mathbb{I}_i^{obj} \|\mathbf{s}_i - \mathbf{E}_{y_i,:}\|_2^2$$

- **Reconstruction Loss**: $$\mathcal{L}_{rec} = \frac{1}{N_p} \sum_{i=1}^{N_p} \|f(x;\Theta) - \mathbf{W}^T \mathbf{s}\|_2^2$$

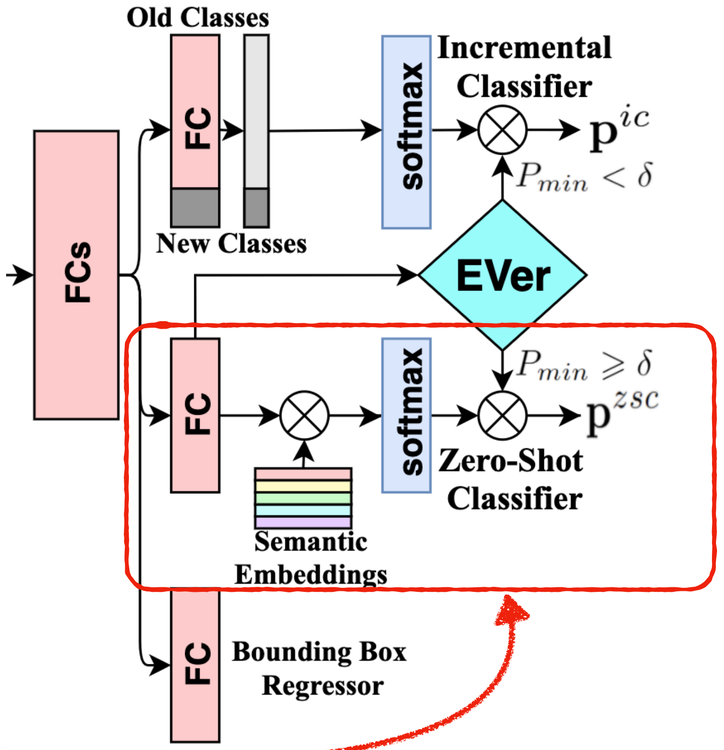
- **Triplet Loss**: $$\mathcal{L}_{tri} = \frac{1}{N_p} \sum_{i=1}^{N_p} \sum_{\substack{j=1 \\ j \neq y_i}}^{N_S} \max(m + C_{ij} - C_{iy_i}, 0)$$

- **overall loss of the backbone**:

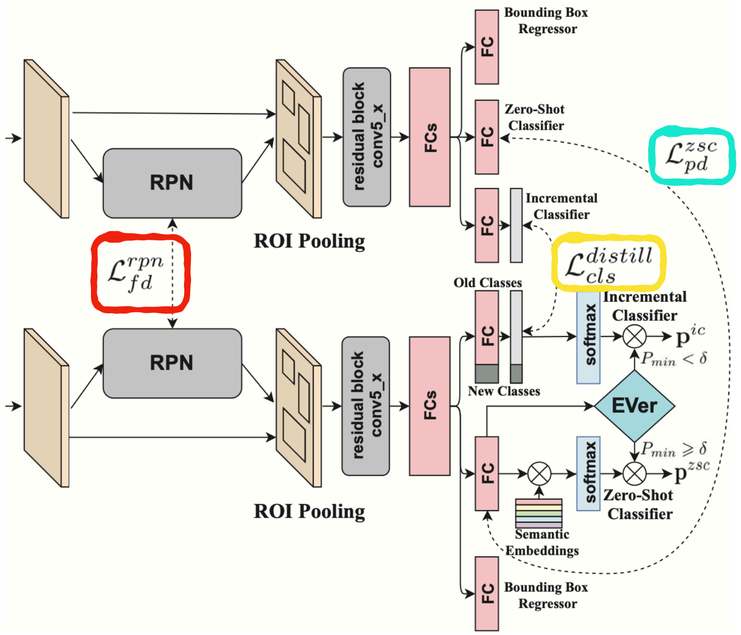
$$\mathcal{L}_{FRCN} = \mathcal{L}_{cls}^{rpn} + \mathcal{L}_{reg}^{rpn} + \mathcal{L}_{reg}$$

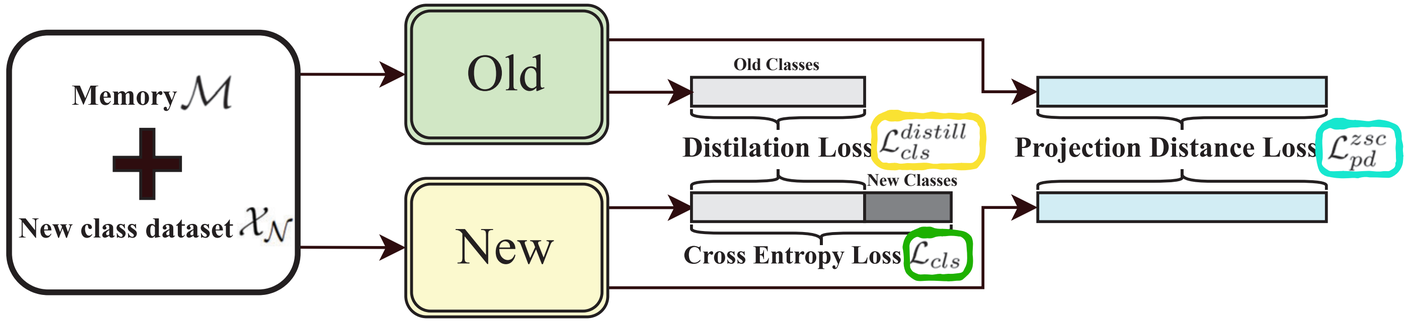
$$\mathcal{L}_{bone} = \mathcal{L}_{FRCN} + \mathcal{L}_{bfmse} + \beta \mathcal{L}_{rec} + \mathcal{L}_{tri}$$

$$\mathcal{L} = \mathcal{L}_{bone} + \mathcal{L}_{cls}$$



# The Old-New Model




- **Feature Distance Loss:** $\mathcal{L}_{fd}^{rpn} = \|f_{rpn}(x;\Theta_o) - f_{rpn}(x;\Theta_n)\|_F^2$
- **Projection Distance Loss:** $\mathcal{L}_{pd}^{zsc} = \|\mathbf{s}_o - \mathbf{s}_n\|_2^2$
- **Distillation Loss:** $\mathcal{L}_{cls}^{distill} = \texttt{CE}(\texttt{softmax}(\frac{\hat{y}_o}{T}), \texttt{softmax}(\frac{\hat{y}_n^o}{T}))$
- **Cross-Entropy Loss:** $\mathcal{L}_{cls} = \texttt{CE}(y, \texttt{softmax}(\hat{y}_n))$
- **overall loss of IZSD-EVer:** $\mathcal{L}_{IL} = \frac{N_{\mathcal{O}}}{N_{\mathcal{C}}}(\mathcal{L}_{cls}^{distill} + \mathcal{L}_{pd}^{zsc}) + \frac{N_{\mathcal{N}}}{N_{\mathcal{C}}}\mathcal{L}_{cls} + \gamma\mathcal{L}_{fd}^{rpn} \qquad \mathcal{L} = \mathcal{L}_{bone} + \mathcal{L}_{IL}$
- **Memory M**: to alleviate the imbalance between old seen and new seen classes

# Extreme Value Analyzer

- Construct EVer based on Pickands-Balkema-de Haan Theorem [1] to differentiate unseen from seen object classes

- use the GPD to model the projected semantic vectors for each seen class

- G(.): probability of being an extreme semantic vector for the corresponding seen classes

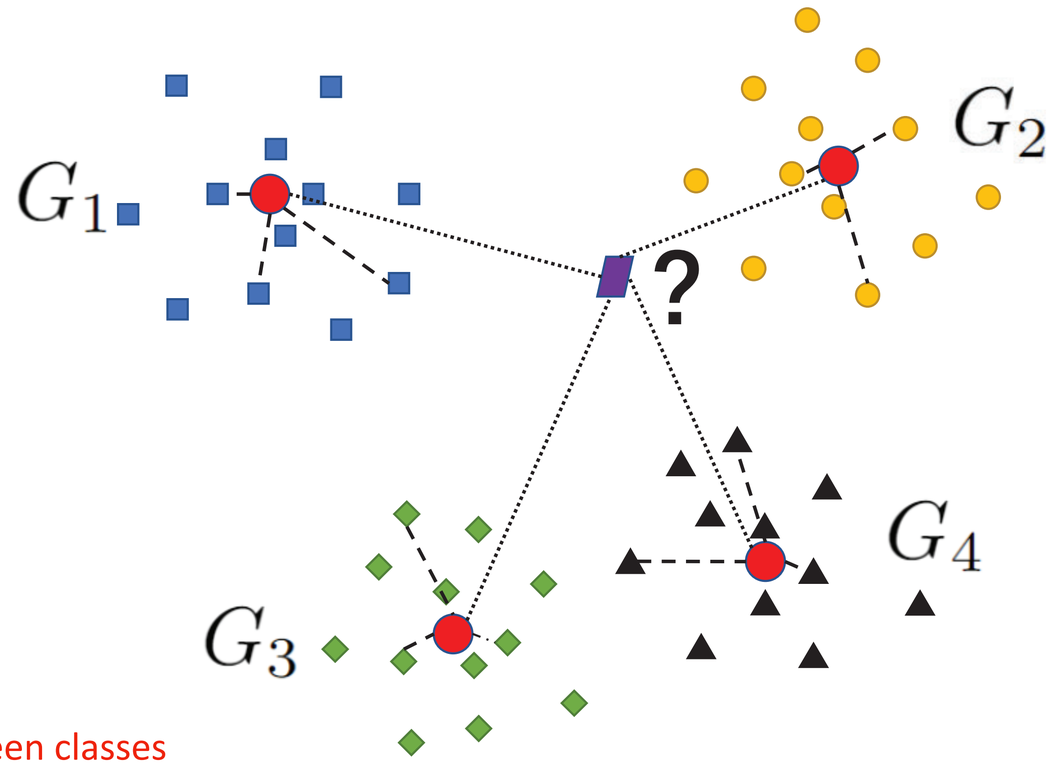
$$P_{min} = \min_{j \in \{1,2,\ldots,S\}} G(\|\mathbf{s} - \bar{\mathbf{s}}_j\|_2 - u_j, \hat{\sigma}_j, \hat{\xi}_j)$$

- The prediction class label：

$$\hat{y} = \begin{cases} \arg\max_{i \in \{1,2,\ldots,S\}} \mathbf{p}_i^{ic}, & P_{min} < \delta \\ \arg\max_{i \in \{1,2,\ldots,U\}} \mathbf{p}_i^{zsc}, & P_{min} \geqslant \delta \end{cases}$$

seen classes

unseen classes




[1] A. A. Balkema and L. De Haan, “Residual life time at great age,” The Annals of probability, pp. 792–804, 1974.

# Experiment Results

Old classe, new classes and Unseen classes split in different incremental step.

| Step | Old classes | New classes | Unseen classes | Train data | Test data |
|---|---|---|---|---|---|
| 1 | - | $\mathcal{G}_1$ | $\mathcal{G}_2, \mathcal{G}_3, \mathcal{G}_4$ | $\mathcal{D}_{tr}(\mathcal{G}_1)$ | $\mathcal{D}_{te}(\mathcal{C})$ |
| 2 | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3, \mathcal{G}_4$ | $\mathcal{D}_{tr}(\mathcal{G}_2)$ | |
| 3 | $\mathcal{G}_1, \mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | $\mathcal{D}_{tr}(\mathcal{G}_3)$ | |
| 4 | $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ | $\mathcal{G}_4$ | - | $\mathcal{D}_{tr}(\mathcal{G}_4)$ | |

# Experiment Results

The results of class-incremental detection on Pascal VOC 2007 and MSCOCO 2017. Gray region means that the results of these groups can only be predicted in our model.

| Dataset | Method | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP | Method | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP | Method | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VOC 2007 | Faster RCNN (Fine-tuning) | 61.9 | - | - | - | 61.9 | CIFRCN | 63.9 | - | - | - | 63.9 | IDWCF | 66.3 | - | - | - | 66.3 |
| | | 33.3 | 67.8 | - | - | 50.6 | | 43.8 | 71.2 | - | - | 57.5 | | 44.8 | 59.2 | - | - | 52.0 |
| | | 13.5 | 23.1 | 55.8 | - | 30.8 | | 35.3 | 49.0 | 68.4 | - | 50.9 | | 49.0 | 43.4 | 48.6 | - | 47.0 |
| | | 14.2 | 21.5 | 44.0 | 52.5 | 33.1 | | 34.6 | 44.1 | 55.6 | 59.6 | 48.5 | | 40.2 | 35.5 | 42.5 | 38.8 | 39.3 |
| | Faster RCNN (Retrain) | 64.0 | - | - | - | 64.0 | CIFRCN NP | 56.8 | - | - | - | 56.8 | Ours | 45.5 | 1.7 | 5.5 | 2.1 | 13.7 |
| | | 65.2 | 73.2 | - | - | 69.2 | | 40.4 | 68.8 | - | - | 54.6 | | 43.0 | 53.4 | 6.5 | 5.2 | 27.0 |
| | | 66.2 | 71.3 | 77.3 | - | 71.6 | | 38.5 | 50.9 | 69.5 | - | 53.0 | | 42.8 | 48.8 | 67.4 | 5.6 | 41.1 |
| | | 65.1 | 70.4 | 76.4 | 67.8 | 69.9 | | 33.7 | 44.6 | 56.8 | 51.8 | 46.7 | | 52.8 | 50.7 | 63.1 | 57.9 | 56.1 |
| COCO 2017 | Faster RCNN (Fine-tuning) | 58.3 | - | - | - | 58.3 | CIFRCN | 58.1 | - | - | - | 58.1 | IDWCF | 49.2 | - | - | - | 49.2 |
| | | 33.5 | 21.0 | - | - | 27.3 | | 39.0 | 24.6 | - | - | 31.8 | | 44.7 | 15.7 | - | - | 30.2 |
| | | 20.1 | 11.9 | 26.4 | - | 19.5 | | 29.9 | 18.5 | 29.6 | - | 26.0 | | 42.4 | 13.1 | 14.0 | - | 23.2 |
| | | 16.8 | 9.3 | 9.4 | 31.4 | 16.7 | | 29.0 | 14.5 | 14.6 | 33.3 | 22.9 | | 38.0 | 12.6 | 11.9 | 21.0 | 20.9 |
| | Faster RCNN (Retrain) | 58.4 | - | - | - | 58.4 | CIFRCN NP | 55.6 | - | - | - | 55.6 | Ours | 46.5 | 2.9 | 0.1 | 0.2 | 12.4 |
| | | 56.1 | 27.4 | - | - | 41.8 | | 38.3 | 24.1 | - | - | 31.2 | | 18.8 | 16.9 | 0.1 | 0.2 | 9.0 |
| | | 54.9 | 26.6 | 30.8 | - | 37.4 | | 26.2 | 18.6 | 27.2 | - | 24.0 | | 10.8 | 10.1 | 20.3 | 0.1 | 10.3 |
| | | 54.8 | 26.3 | 28.0 | 37.2 | 36.6 | | 26.5 | 15.1 | 15.5 | 32.2 | 22.3 | | 38.4 | 15.9 | 12.3 | 26.9 | 23.4 |

# Experiment Results

- Ablation Study

The Effect of bfMSE

| Step | bfMSE | | | | | MSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP |
| 1 | 45.45 | 1.65 | 5.47 | 2.10 | 13.67 | 38.88 | 1.72 | 4.15 | 1.71 | 11.61 |
| 2 | 42.96 | 53.42 | 6.49 | 5.22 | 27.02 | 28.12 | 51.71 | 5.98 | 5.67 | 22.87 |
| 3 | 42.82 | 48.78 | 67.38 | 5.59 | 41.14 | 26.86 | 44.71 | 58.31 | 2.51 | 33.10 |
| 4 | 52.79 | 50.73 | 63.05 | 57.86 | 56.11 | 42.99 | 46.27 | 52.56 | 50.00 | 47.96 |

The effect of PD loss

| Step | $\mathcal{L}_{pd}^{zsc}$ | | | | | without $\mathcal{L}_{pd}^{zsc}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP | $\mathcal{G}_1$ | $\mathcal{G}_2$ | $\mathcal{G}_3$ | $\mathcal{G}_4$ | mAP |
| 1 | 45.45 | 1.65 | 5.47 | 2.10 | 13.67 | 36.06 | 1.8 | 2.54 | 1.62 | 10.5 |
| 2 | 42.96 | 53.42 | 6.49 | 5.22 | 27.02 | 29.99 | 50.59 | 6.81 | 5.08 | 23.12 |
| 3 | 42.82 | 48.78 | 67.38 | 5.59 | 41.14 | 25.16 | 44.89 | 58.6 | 2.85 | 32.88 |
| 4 | 52.79 | 50.73 | 63.05 | 57.86 | 56.11 | 42.86 | 48.43 | 56.21 | 50.05 | 49.39 |



# Thanks