# Contrastive Prototypical Network with Wasserstein Confidence Penalty


Haoqing Wang, Zhi-Hong Deng

School of Intelligence Science and Technology, Peking University


## 1. Background

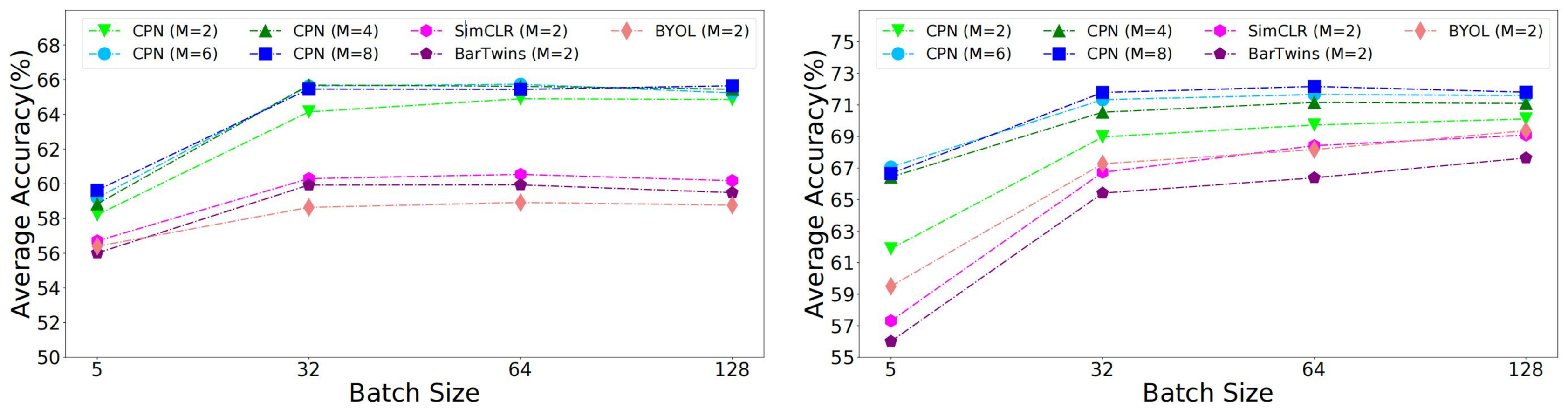
### Unsupervised Few-Shot Learning

- Few-shot learning
  - it is hard for machine to solve a *novel* task based on limited labeled data.
  - one can learn task-shared inductive bias from a base dataset beforehand.
- Unlabeled base dataset
  - obtaining sufficient labeled data for certain domains may be difficult or even impossible in practice, such as satellite imagery and skin diseases.
  - **learn the inductive bias in the unsupervised manner**
- Sampling-Augmentation paradigm: given an unlabeled dataset $\mathcal{D}$, samples $\{x_i\}_{i=1}^N$ are randomly selected and each $x_i$ represents a pseudo class. For each $x_i$, in-class mples $\{v_i^j\}_{j=1}^M$ are generated via manually or learnable data augmentations. For a specific problem, the loss function $\mathcal{L}$ is calculated on the sub-dataset $\{v_i^j\}_{i=1,j=1}^{N,M}$ and the training objective is

$$\min_{\theta} \mathbb{E}_{p(\{v_i^j\}_{i=1,j=1}^{N,M})} \left[\mathcal{L}\left(\{v_i^j\}_{i=1,j=1}^{N,M}, \theta\right)\right]$$

where $\theta$ represents the model parameter.

- *data augmentation based unsupervised few-shot learning models*
  - Set $M = S + Q$, we get $N \bullet S$ support samples and $N \bullet Q$ query samples
  - small $N$ (e.g., 5) and big $M$ (e.g., 5+15)
- *contrastive learning models*
  - InfoNCE loss
  - huge $N$ (e.g., 4096) and tiny $M$ (e.g., 2)




(a) Backbone: Conv4 (b) Backbone: ResNet12

Figure 1: Average few-shot classification accuracy across four different settings (5-way 1-shot/5-shot/20-shot/50-shot) on miniImageNet with varying batch size. Here 'CPN' represents prototypical loss with pairwise contrast.

## 2. Contrastive Prototypical Network

- Empirical study on Sampling-Augmentation paradigm (Fig. 1)
  - the loss function $\mathcal{L}$, the batch size $N$ and the view number $M$
  - in the few-shot learning, with the same batch size the contrastive losses perform worse than the prototypical loss which directly compares the representations of different views.
  - unsupervised few-shot learning prefers large batch size.
  - more augmented views lead to better performance due to increased view diversity.
- Pairwise contrast is useful.

| Model | Backbone | 1-shot | 5-shot | 20-shot | 50-shot |
|---|---|---|---|---|---|
| CPN w/o PC | Conv4 | 46.08 ± 0.19 | 63.89 ± 0.17 | 72.59 ± 0.14 | 74.81 ± 0.14 |
| CPN | | 46.96 ± 0.19 | 64.75 ± 0.17 | 73.31 ± 0.14 | 75.63 ± 0.14 |
| CPN w/o PC | ResNet12 | 48.80 ± 0.19 | 69.09 ± 0.16 | 78.54 ± 0.13 | 80.83 ± 0.12 |
| CPN | | 50.01 ± 0.18 | 70.73 ± 0.16 | 80.33 ± 0.13 | 82.74 ± 0.11 |

- Contrastive Prototypical Network
  - the $l$-th view $\{v_i^l\}_{i=1}^N$ is used as the one-shot support set to classify all the views.
  - prototypical loss with pairwise contrast and large batch size.

## 3. Wasserstein Confidence Penalty

- Sampling bias
  - some negative pairs (e.g., $(a_1, b_2)$) may be semantically similar or even belong to the same semantic class.
  - using the one-hot prediction target could overly push the semantically similar negative pairs away from each other and has the risk of learning sample-specific information.

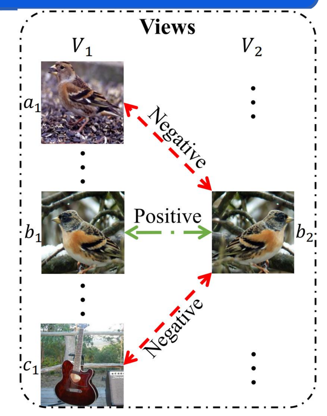



Figure 2. Sampling bias.

- Penalizing over-confident prediction
  - making the prediction $\mathrm{p}$ approximating a latent distribution $\mathrm{q}$ (i.e., the uniform distribution)
  - existing regularization methods based on f-divergence: $\mathrm{D}(\mathrm{p}, \mathrm{q}) = \sum_{k=1}^{N} f(p_k/q_k)$
    - Label Smoothing: $f(z) = -\ln z$
    - Confidence Penalty: $f(z) = z \ln z$
  - Wasserstein Confidence Penalty
    - the difference in the probability of each class is computed independently in f-divergence and the structural information, i.e., the semantic relationships among different classes, is ignored.
    - we use the Wasserstein distance as $\mathrm{D}(\mathrm{p}, \mathrm{q})$ and introduce the structural information using the cost matrix.
    - the transportation cost between pseudo class $i$ and pseudo class $j$

$$C_{ij} = \gamma \bullet (1 - S_{ij}) + \mathbb{I}_{i=j}$$

where $\gamma$ is a scaling factor, $S_{ij}$ represents the semantic similarity between class $i$ and class $j$, $\mathbb{I}_{ij}$ is an indicator function in the condition $i = j$.

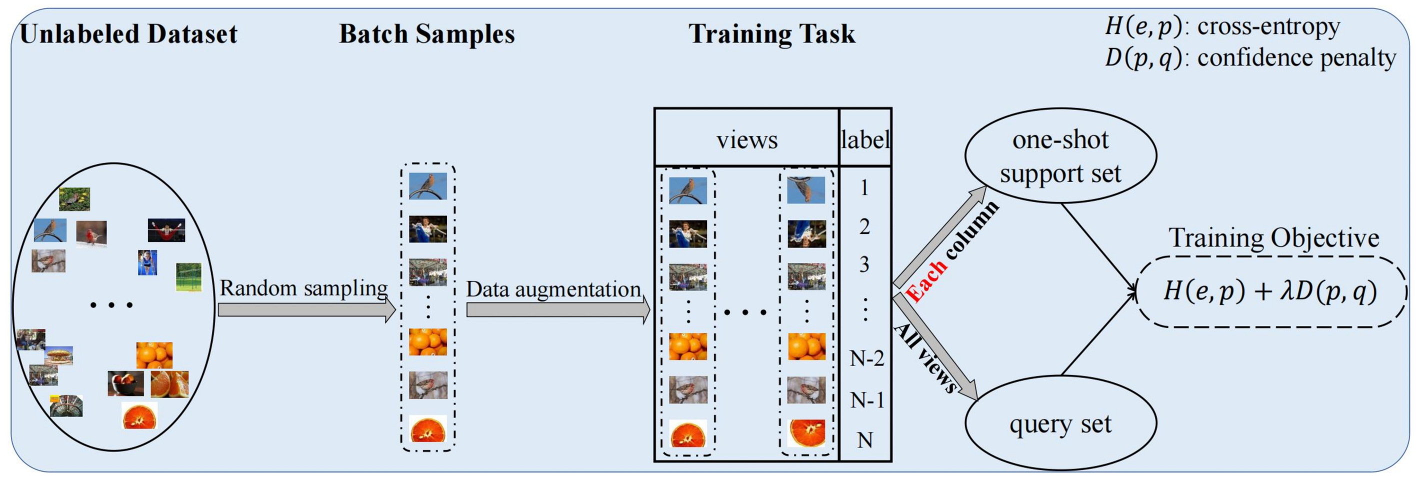



Figure 3.CPNWCP

## 4. Experiments

### A. SOTA in unsupervised few-shot learning

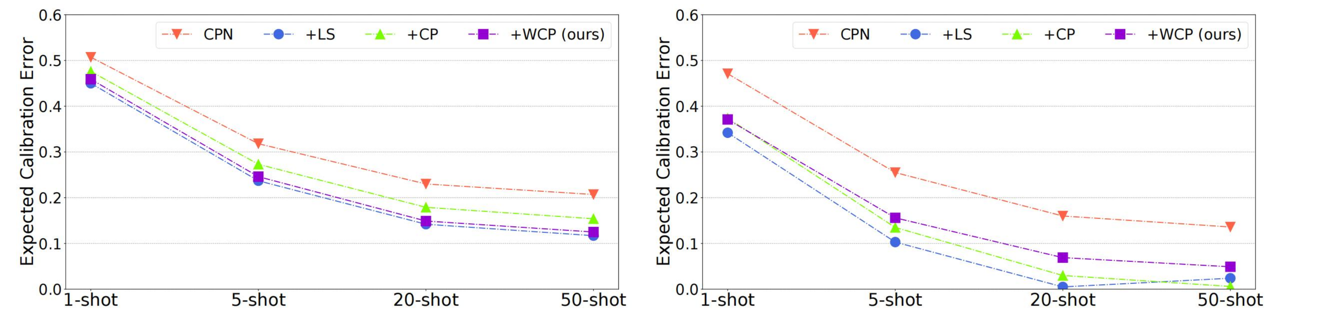
| Model | 1-shot | 5-shot | 20-shot | 50-shot |
|---|---|---|---|---|
| Train from scratch [21] | 27.59 ± 0.59 | 38.48 ± 0.66 | 51.53 ± 0.72 | 59.63 ± 0.74 |
| CACTUs-ProtoNet [21] | 39.18 ± 0.71 | 53.36 ± 0.70 | 61.54 ± 0.68 | 63.55 ± 0.64 |
| CACTUs-MAML [21] | 39.90 ± 0.74 | 53.97 ± 0.70 | 63.84 ± 0.70 | 69.64 ± 0.63 |
| UMTRA [25] | 39.93 | 50.73 | 61.11 | 67.15 |
| ULDA-ProtoNet [36] | 40.63 ± 0.61 | 56.18 ± 0.59 | 64.31 ± 0.51 | 66.43 ± 0.47 |
| ULDA-MetaOptNet [36] | 40.71 ± 0.62 | 54.49 ± 0.58 | 63.58 ± 0.51 | 67.65 ± 0.48 |
| LASIUM-ProtoNet [26] | 40.05 ± 0.60 | 52.53 ± 0.51 | 59.45 ± 0.48 | 61.43 ± 0.45 |
| LASIUM-MAML [26] | 40.19 ± 0.58 | 54.56 ± 0.55 | 65.17 ± 0.49 | 69.13 ± 0.49 |
| ArL-RelationNet [54] | 36.37 ± 0.92 | 46.97 ± 0.86 | - | - |
| ArL-ProtoNet [54] | 38.76 ± 0.84 | 51.08 ± 0.84 | - | - |
| ArL-SoSN [54] | 41.13 ± 0.84 | 55.39 ± 0.79 | - | - |
| SimCLR [9] | 40.91 ± 0.19 | 57.22 ± 0.17 | 65.74 ± 0.15 | 67.83 ± 0.15 |
| BYOL [16] | 39.81 ± 0.18 | 56.65 ± 0.17 | 64.58 ± 0.15 | 66.69 ± 0.15 |
| BarTwins [51] | 39.02 ± 0.18 | 57.20 ± 0.17 | 65.26 ± 0.15 | 67.42 ± 0.14 |
| ProtoCLR [30] | 44.89 ± 0.58 | 63.35 ± 0.54 | 72.27 ± 0.45 | 74.31 ± 0.45 |
| CPNWCP (ours) | **47.93 ± 0.19** | **66.44 ± 0.17** | **75.69 ± 0.14** | **78.20 ± 0.13** |
| ProtoNet-Sup [40] | 49.42 ± 0.78 | 68.20 ± 0.66 | - | - |

### B. Analytical experiments

- Wasserstein Confidence Penalty can more effectively alleviate the sampling bias.

| Model | Backbone | 1-shot | 5-shot | 20-shot | 50-shot |
|---|---|---|---|---|---|
| CPN | Conv4 | 46.96 ± 0.19 | 64.75 ± 0.17 | 73.31 ± 0.14 | 75.63 ± 0.14 |
| + CR [48] | | 47.33 ± 0.19 | 65.15 ± 0.17 | 73.28 ± 0.14 | 75.50 ± 0.14 |
| + LS [41] | | 47.19 ± 0.19 | 65.22 ± 0.17 | 74.21 ± 0.14 | 76.71 ± 0.13 |
| + CP [34] | | 47.22 ± 0.19 | 65.46 ± 0.17 | 74.52 ± 0.14 | 77.05 ± 0.13 |
| + JSCP | | 46.82 ± 0.19 | 64.89 ± 0.17 | 73.92 ± 0.14 | 76.37 ± 0.13 |
| + WCP (ours) | | **47.93 ± 0.19** | **66.44 ± 0.17** | **75.69 ± 0.14** | **78.20 ± 0.13** |
| CPN | ResNet12 | 50.01 ± 0.18 | 70.73 ± 0.16 | 80.33 ± 0.13 | 82.74 ± 0.11 |
| + CR [48] | | 51.85 ± 0.19 | 72.23 ± 0.16 | 81.35 ± 0.12 | 83.28 ± 0.11 |
| + LS [41] | | 50.41 ± 0.19 | 71.10 ± 0.16 | 80.97 ± 0.12 | 83.61 ± 0.11 |
| + CP [34] | | 50.71 ± 0.18 | 71.29 ± 0.16 | 81.11 ± 0.12 | 83.91 ± 0.11 |
| + JSCP | | 49.87 ± 0.18 | 70.53 ± 0.16 | 81.01 ± 0.13 | 83.19 ± 0.11 |
| + WCP (ours) | | **53.56 ± 0.19** | **73.21 ± 0.16** | **82.18 ± 0.12** | **84.35 ± 0.11** |

- Wasserstein Confidence Penalty can improve the prediction calibration.




(a) Backbone: Conv4 (b) Backbone: ResNet12