MICM: Rethinking Unsupervised Pretraining for Enhanced Few-shot Learning

Supplementary Materials

1 MODEL ROBUSTNESS

1.1 Further Cross-Domain Few-Shot Comparing

In addition, cross-domain experiments were conducted on the CUB dataset [11], characterized by a relatively minor domain gap. Adhering to the protocols of Poulakakis et al. [10], we trained MICM on the miniImageNet and evaluated it on the CUB dataset's test set for both 5-way 1-shot and 5-way 5-shot classification tasks. We present the performance results of our MICM model compared to existing unsupervised methods.

As reported in Table 1, our model not only surpasses existing unsupervised methods but also achieves a significant improvement of 4 / 4.3 points over the SOTA Transductive U-FSL method BECLR [10] in 1-shot and 5-shot tasks, respectively.

Table 1: Accuracies (in % ± std) on miniImageNet → CUB. The results of the existing model are cited from BECLR [10]

Method	miniImageNet → CUB	
	5-way 1-shot	5-way 5-shot
Meta-GMVAE [9]	38.04±0.47	55.65±0.42
SimCLR [3]	38.25±0.49	55.89±0.46
MoCo v2 [7]	39.29±0.47	56.49±0.44
BYOL [6]	40.63±0.46	56.92±0.43
SwAV [1]	38.34±0.51	53.94±0.43
NNCLR [5]	39.37±0.53	54.78±0.42
Barlow Twins [13]	40.46±0.47	57.16±0.42
Laplacian Eigenmaps [2]	41.08±0.48	58.86±0.45
HMS [12]	40.75	58.32
PsCo [8]	-	57.38±0.44
BECLR [10]	43.45±0.50	59.51±0.46
MICM **(OURS)**	**47.44**±**0.65**	**63.86**±**0.42**

1.2 Sample Bias

Sample bias is an important factor that influences few-shot learning. To evaluate the robustness of our method against sample bias, we investigate two strategies for its mitigation: 1) augmenting the number of support samples, and 2) refining the class prototype using an increased number of query samples (Here, we choose to refine the class prototype using the OpTA algorithm [10]). To assess the effectiveness of these strategies, we analyze performance variations of both our model and the baseline across different N-way, K-shot, Q-query configurations. This analysis involves, as illustrated in Figure 1, incrementally increasing 1) the number of support samples (K) and 2) the number of query samples (Q). From these experiments, a clear trend emerges: as the count of support or query samples rises – effectively reducing sample bias – the superiority of our model over the baseline becomes increasingly evident. This observation underscores the enhanced adaptability of our approach, especially in scenarios characterized by smaller sample bias, where our model demonstrates a more substantial performance improvement compared to the baseline.

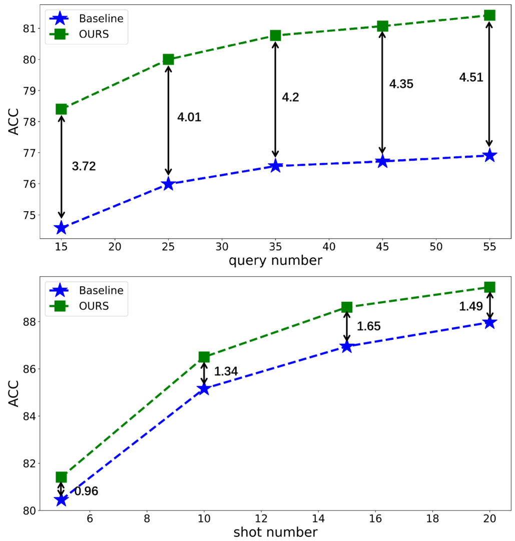

Figure 1: Performance comparison for varying numbers of shots and queries.

2 MICM+

We further developed a hybrid method, **MICM+**, by integrating pseudo-label learning [10] with the transductive OpTA FSL technique [10]. This approach exploits the synergistic potentials of both methods to enhance performance in scenarios with limited labeled data.

2.1 Pseudo Label Training

BECLR's Pseudo Label Training Stage. The SOTA BECLR [10] utilizes a memory bank alongside clustering techniques to facilitate pseudo-label training. Despite its effectiveness, this method faces challenges such as increased storage requirements and slow convergence rates, stemming from continuous updates between the memory bank and current samples during training.

Table 2: Our MICM model's inductive few-shot performance on the MiniImageNet dataset after incorporating two pseudo-label training methods.

Method	Training time	5-way 1-shot	5-way 5-shot
MICM	31.25 Hours	61.37±0.62	81.68±0.43
MICM + BECLR (From scratch)	46.25 Hours	57.43±0.62	77.34±0.51
MICM + BECLR (A new stage)	31.25 + 1.50 Hours	61.30±0.59	81.65±0.37
MICM + Pseudo (A new stage)	31.25 + 0.16 Hours	66.69±0.65	84.03±0.45

In our architecture, we experimented with integrating BECLR's pseudo-label training either from scratch or into a pre-trained MICM model. Our findings, detailed in Table 2, reveal that starting from scratch prolongs training times and diminishes performance, as does introducing pseudo-labeling to a pre-trained model. To address these issues, we propose a novel pseudo-label training strategy using BCE loss, as shown in Figure 2. This method avoids additional storage costs and can be seamlessly added to existing pre-trained models.

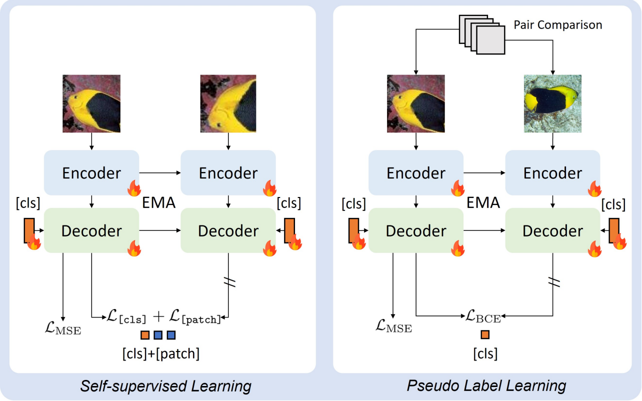

Figure 2: Our pseudo label training stage: further refines samples representations using pair comparison techniques, thereby enhancing the model's ability to differentiate between various image representations.

Our Pseudo Label Training Stage. Building upon the robust representations developed during the self-supervised learning stage, our primary objective is to enhance the inter-class distinction. To achieve this, we have incorporated a pseudo-label learning method aimed at increasing intra-class compactness. This approach is detailed in Figure 2 and employs a pairwise objective to promote similarity between instance pairs, ensuring effective clustering of instances within the same class.

Pseudo-labels are generated by calculating the cosine distances among all pairs of feature representations, $\mathbf{Z}_{\texttt{[cls]}}$, within a mini-batch. These distances are ranked, and each instance is assigned a pseudo-label based on its closest neighbor. Pseudo-labels are thus generated from the most confidently paired positive instances in the mini-batch. Given a mini-batch, S, containing B instances with their features $\mathbf{Z}_{\texttt{[cls]}}$, we denote the subset of closest pairs as S′. The pairwise objective is defined using a binary cross-entropy loss (BCE) as follows:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{B}\sum_{i=1}^{B} -\log\langle\sigma(\mathbf{Z}^{(i)}_{\texttt{[cls]}}), \sigma(\mathbf{Z}'^{(i)}_{\texttt{[cls]}})\rangle \tag{1}$$

where σ is a normalization function applied to each feature vector in S and S′.

In addition to the BCE loss, we continue to use the mean squared error (MSE) loss, $\mathcal{L}_{\text{MSE}}$, from the self-supervised stage to maintain a balance between classification efficacy and model generalization.

Furthermore, we have opted to remove the patch-level loss, $\mathcal{L}_{\texttt{[patch]}}$, for two primary reasons: Firstly, our pseudo-label training does not involve comparing two views of the same image, making patch-level alignment infeasible. Secondly, maintaining two views and computing patch loss would significantly increase storage demands, necessitating a reduction in batch size. Larger batch size is essential for effective pseudo-label training. This modification, as documented in Table 3, involved reducing the batch size from 128 to 80, leading to a degradation in model performance.

Table 3: MICM+'s transductive few-shot performance after pseudo-label training with/without $\mathcal{L}_{\texttt{[patch]}}$. When $\mathcal{L}_{\texttt{[patch]}}$ is retained, a smaller batch size is required due to the increased GPU memory consumption.

Method	5-way 1-shot	5-way 5-shot
MICM+ w/ Patch loss	77.96±0.65	85.46±0.41
MICM+	**81.05±0.58**	**87.95±0.34**

2.2 Optimal Transport-based Distribution Alignment (OpTA)

In line with BECLR's task setting [10], we employ the OpTA algorithm for transductive few-shot classification tasks. The OpTA process is expressed as follows:

Let $\mathcal{T} = \mathcal{S} \cup \mathcal{Q}$ be a downstream few-shot task. We first extract the support $Z^{\mathcal{S}}$ (of size NK × d) and query $Z^{\mathcal{Q}}$ (of size NQ × d) embeddings and calculate the support set prototypes $\boldsymbol{P}^{\mathcal{S}}$ (class averages of size N × d). Firstly, an optimal transport problem is defined from $Z^{\mathcal{Q}}$ to $\boldsymbol{P}^{\mathcal{S}}$ as:

$$\Pi(\boldsymbol{r},\boldsymbol{c}) = \left\{\boldsymbol{\pi} \in \mathbb{R}_{+}^{NQ\times N} \mid \boldsymbol{\pi}\mathbf{1}_N = \boldsymbol{r}, \boldsymbol{\pi}^{\top}\mathbf{1}_{NQ} = \boldsymbol{c}, \boldsymbol{r} = \mathbf{1}\cdot 1/NQ, \boldsymbol{c} = \mathbf{1}\cdot 1/N\right\} \tag{2}$$

To find a transport plan π (out of Π) mapping $Z^{\mathcal{Q}}$ to $\boldsymbol{P}^{\mathcal{S}}$. Here, $\boldsymbol{r} \in \mathbb{R}^{NQ}$ denotes the distribution of batch embeddings $[\boldsymbol{z}_i]_{i=1}^{NQ}$, $\boldsymbol{c} \in \mathbb{R}^{N}$ is the distribution of prototypes $[\boldsymbol{P}_i]_{i=1}^{N}$. The last two conditions in Eq. 2 enforce equipartitioning (i.e., uniform assignment) of Z into the P partitions. Obtaining the optimal transport plan, $\hat{\boldsymbol{\pi}}^{\star}$, can then be formulated as:

$$\boldsymbol{\pi}^{\star} = \underset{\boldsymbol{\pi}\in\Pi(\boldsymbol{r},\boldsymbol{c})}{\operatorname{argmin}} \langle\boldsymbol{\pi}, \boldsymbol{D}\rangle_F - \varepsilon\mathbb{H}(\boldsymbol{\pi}), \tag{3}$$

and solved using the Sinkhorn-Knopp [4] algorithm. Here, D is a pairwise distance matrix between the elements of $Z^{\mathcal{Q}}$ and $\boldsymbol{P}^{\mathcal{S}}$ (of size NQ × N), $\langle\cdot\rangle_F$ denotes the Frobenius dot product, ε is a regularisation term, and $\mathbb{H}(\cdot)$ is the Shannon entropy.

After Obtaining the optimal transport plan $\hat{\boldsymbol{\pi}}^{\star}$, we use $\hat{\boldsymbol{\pi}}^{\star}$ to map the support set prototypes onto the region occupied by the query embeddings to get the transported support prototypes $\hat{\boldsymbol{P}}^{\mathcal{S}}$ as:

$$\hat{\boldsymbol{P}}^{\mathcal{S}} = \hat{\boldsymbol{\pi}}^{\star T} Z^{\mathcal{Q}}, \quad \hat{\boldsymbol{\pi}}^{\star}_{i,j} = \frac{\boldsymbol{\pi}^{\star}_{i,j}}{\sum_j \boldsymbol{\pi}^{\star}_{i,j}}, \forall i \in [NQ], j \in [N], \tag{4}$$

and a comprehensive description of this algorithm is provided in BECLR [10]. Our application of MICM+ with OpTA has led to improved transductive few-shot performance, discussed in subsequent sections.

Table 4: Ablating main components of MICM.

OpTA	MICM	$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{CL}}$	Pseudo	5-way 1-shot	5-way 5-shot
-	-	-	-	-	60.93 ± 0.61	80.38 ± 0.34
✓	-	-	-	-	74.58 ± 0.66	83.95 ± 0.34
✓	✓	✓	-	-	26.81 ± 0.43	32.94 ± 0.47
✓	✓	-	✓	-	75.73 ± 0.64	85.06 ± 0.35
✓	✓	✓	✓	-	78.40 ± 0.61	86.90 ± 0.30
✓	✓	✓	✓	✓	**81.05 ± 0.58**	**87.95 ± 0.34**

3 ABLATION STUDY

The proposed MICM+ model integrates five key components incuding: OpTA [10], MICM, ($\mathcal{L}_{\text{MSE}}$), ($\mathcal{L}_{\text{CL}}$) and pseudo-label learning (pseudo). As detailed in Table 4, the baseline model employing OpTA for transductive classification tasks exhibits a notable improvement of 13.6% and 3.5% over traditional inductive classification approaches. This marked enhancement, especially in the 1-shot scenario, can be attributed to OpTA's effective mitigation of sample bias. Our model MICM combines the $\mathcal{L}_{\text{CL}}$ from CL and the $\mathcal{L}_{\text{MSE}}$ from MIM, but in the ablation experiments, we separate these two loss functions. It can be observed that the model using only $\mathcal{L}_{\text{MSE}}$ has no classification ability, while the model using only $\mathcal{L}_{\text{CL}}$ shows relatively good classification performance. However, by combining $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{CL}}$, the model's performance improves by approximately 2.7 / 1.9 points. This result highlights the importance of utilizing generalized features learned during the image reconstruction process. Integration of pseudo-label training contributes additional gains of 2.5% and 1.0% in the 1-shot and 5-shot setups. This enhancement, facilitated by pseudo, further elevate the representational capabilities of features from the pre-training stage and their adaptability to small sample data in downstream tasks.

REFERENCES

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

[2] Kuilin Chen and Chi-Guhn Lee. 2022. Unsupervised Few-shot Learning via Deep Laplacian Eigenmaps. *arXiv preprint arXiv:2210.03595* (2022).

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[4] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).

[5] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9588–9597.

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[8] Huiwon Jang, Hankook Lee, and Jinwoo Shin. 2023. Unsupervised Meta-learning via Few-shot Pseudo-supervised Contrastive Learning. *arXiv preprint arXiv:2303.00996* (2023).

[9] Dong Bok Lee. 2021. Meta-GMVAE: Mixture of Gaussian VAEs for unsupervised meta-learning. (2021).

[10] Stelios Poulakakis Daktylidis. 2023. BECLR: Batch Enhanced Contrastive Unsupervised Few-Shot Learning. (2023).

[11] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD birds 200. (2010).

[12] Han-Jia Ye, Lu Han, and De-Chuan Zhan. 2022. Revisiting unsupervised meta-learning via the characteristics of few-shot tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3721–3737.

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*. PMLR, 12310–12320.