Is Meta-Learning Out? Rethinking Unsupervised Few-Shot Classification with Limited Entropy

hustgyc@hust.edu.cn

Background: Whole class training(WCT) outperform meta-learning in few-shot classification tasks

Question: Is meta-learning still matter?

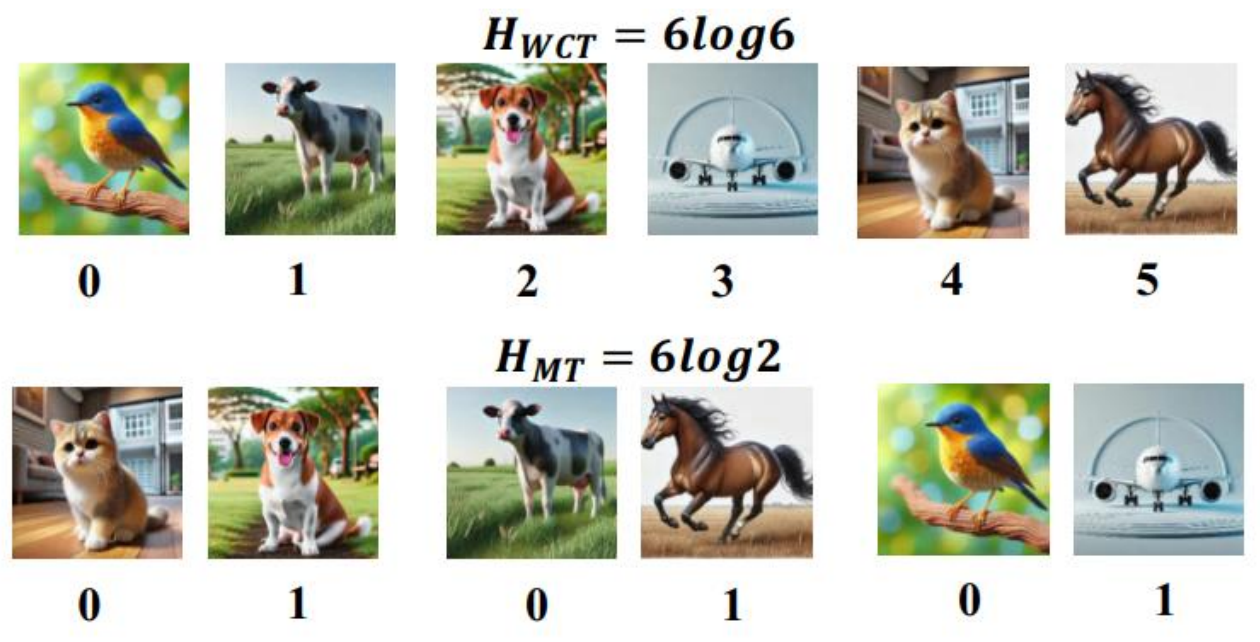


Figure 1. Unfair comparison under the conventional supervised setting. The annotation cost varies across different training methods. H represents the information entropy.

Unfair comparison:

- WCT requires more category distinctions → higher entropy
- Meta-training uses limited task categories → lower entropy

Motivation:
Under limited entropy, does meta-learning still better?

Contribution 1: Entropy-Limited Supervision

Lemma 1. *Let the sample volume of the dataset be m, the number of classes be C, the sample number per class be balanced, and the entropy consumed by annotation be H. Then, the expectation of correct labeled samples, i.e., m', is given by*

$$m' = \frac{m}{C} e^{\frac{H}{m}} \qquad s.t. \;\; H \in [0, mlogC]. \tag{1}$$

Corollary 1. *Let the base-level stability $\beta \sim o(\sqrt{1/m})$, the meta-level stability $\tilde{\beta} \sim o(\sqrt{1/n})$, and the entropy resource H be equal for each algorithm. Then, the meta-learning algorithm $\boldsymbol{\mathcal{A}}$ has a tighter generalization error upper bound than the single-task learning algorithm $\boldsymbol{A}$ when*

$$C_2^2 \cdot k < C_1. \tag{4}$$

Contribution 2: MINO

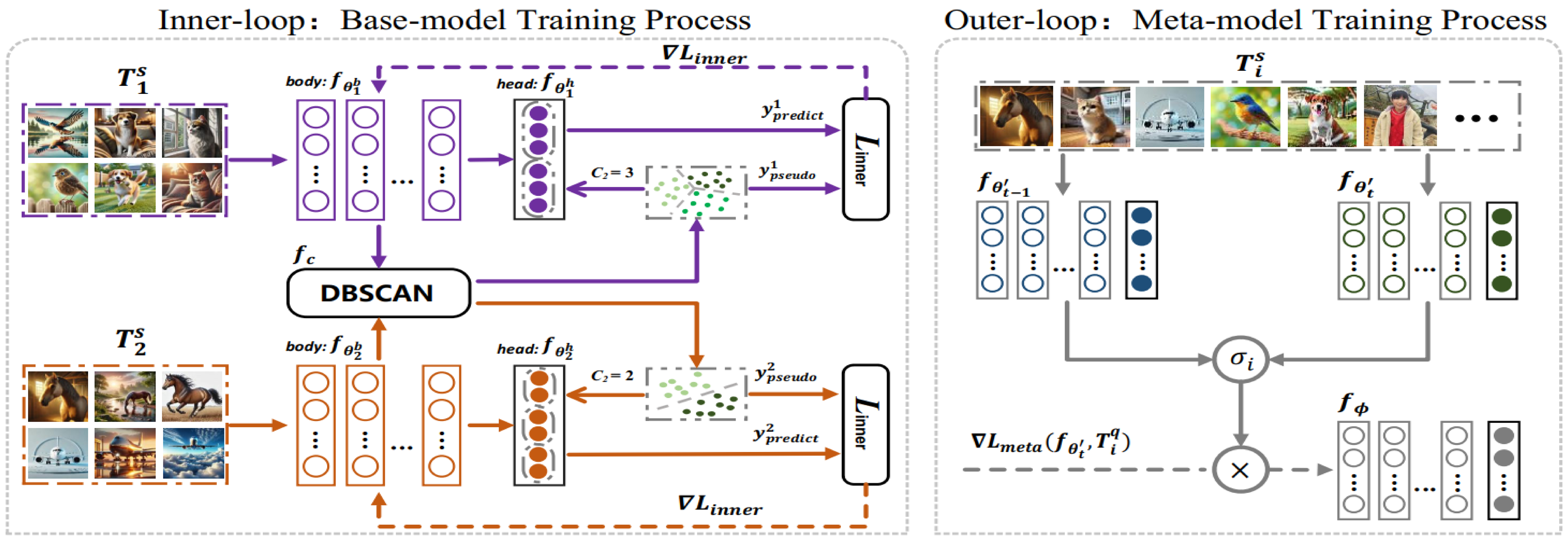
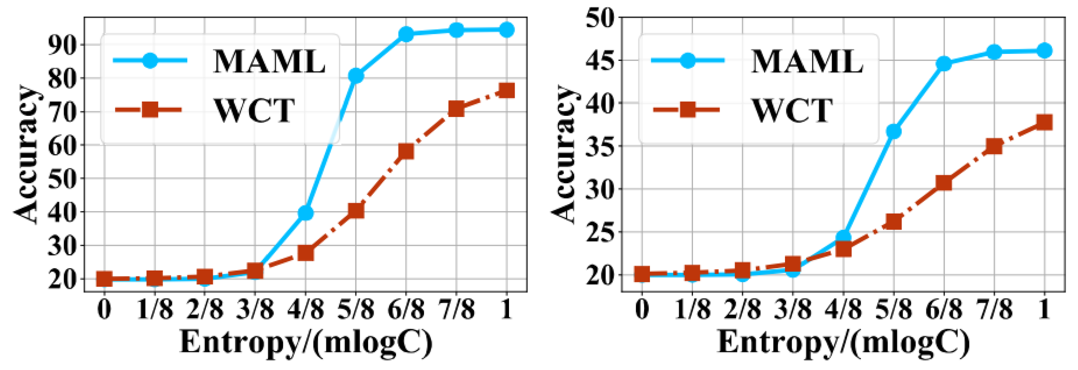


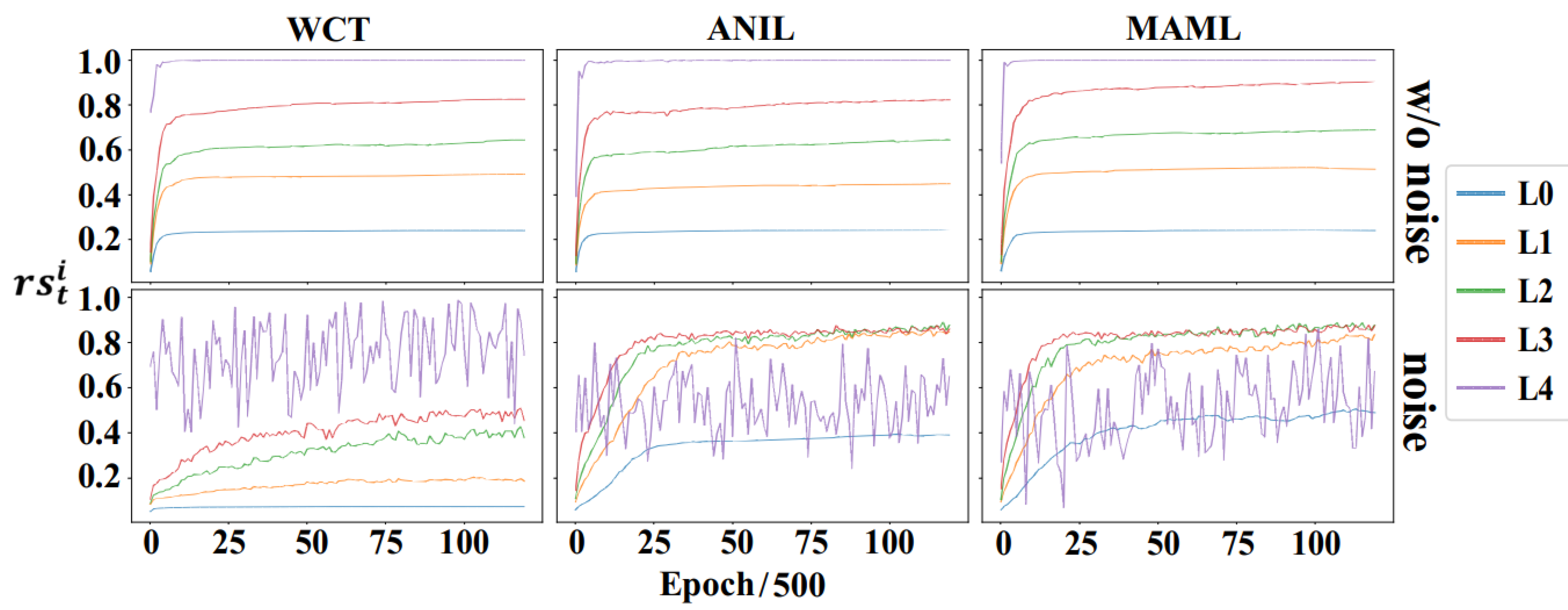
Figure 5. Overview of MINO. (1) The left part shows the process of inner-loop. The base learner computes y_{pseudo} and $y_{predict}$ by DBSCAN f_c and a dynamic head $f_{\theta_i^h}$, respectively. Then, their cross-entropy is backpropagated. We apply a grouping classification technique at the head, *i.e.*, f_{θ^h}, to handle tasks with different numbers of clusters. (2) The right part shows the process of outer-loop. The meta gradient, *i.e.*, $\nabla L_{meta}(f_{\theta'_t}, T_i^q)$, is regulated by the meta-scaler, *i.e.*, σ_i, where σ_i is an adaptive scaler that operates based on the representation stability of $f_{\theta'_t}$ to ensure noise robustness.

Theoretical and Experimental results:

- Entropy-Limited Supervision Bridges the gap between supervised and unsupervised settings.
- Under limited entropy, meta-learning is better.



(a) Omniglot 5-way 1-shot (b) Mini-Imagenet 5-way 1-shot



Entropy efficiency

Noise robustness

Better accuracy

Method	CIFAR-10	CIFAR-100	STL-10	ImageNet	Tiny-MINIST	DomainNet
DeepCluster [2]	63.02 ± 1.14	35.05 ± 1.11	52.21 ± 1.42	24.83 ± 0.95	78.63 ± 1.68	18.09 ± 0.88
IIC [18]	64.05 ± 1.02	36.23 ± 1.27	53.78 ± 1.30	25.07 ± 0.88	79.21 ± 1.54	18.18 ± 0.74
MAE [14]	68.83 ± 1.19	39.11 ± 1.52	56.19 ± 1.47	27.32 ± 1.14	81.03 ± 1.36	20.53 ± 1.03
NVAE [35]	67.43 ± 1.37	38.29 ± 1.45	55.78 ± 1.22	27.21 ± 0.98	81.52 ± 1.61	19.84 ± 0.79
BiGAN [9]	67.61 ± 1.24	38.78 ± 1.19	55.24 ± 1.34	26.85 ± 1.07	80.09 ± 1.27	19.23 ± 0.95
ReSSL [41]	70.27 ± 1.15	41.48 ± 1.60	58.52 ± 1.31	31.25 ± 1.13	83.17 ± 1.24	21.42 ± 0.92
Meta-GMVAE [23]	71.73 ± 1.28	41.26 ± 1.02	58.69 ± 1.51	30.08 ± 1.57	84.65 ± 1.03	21.06 ± 1.37
MINO-kmeans	69.06 ± 1.34	39.55 ± 1.27	57.05 ± 1.49	29.89 ± 1.83	83.15 ± 1.01	19.68 ± 1.16
MINO	73.15 ± 1.09	43.34 ± 1.41	60.74 ± 1.35	31.12 ± 1.06	86.45 ± 1.28	22.68 ± 0.91

	Omniglot				
(way, shot)	(5, 1)	(5, 5)	(20, 1)	(20, 5)	(5, 1)
UMTRA [21]	82.97 ± 0.68	94.84 ± 0.60	73.51 ± 0.53	91.22 ± 0.59	39.14 ± 1.02
CACTUs-MA-DC [15]	67.98 ± 0.80	87.07 ± 0.63	47.48 ± 0.59	72.21 ± 0.54	39.11 ± 1.08
CACTUs-Pr-DC [15]	67.08 ± 0.72	82.97 ± 0.64	46.32 ± 0.51	65.75 ± 0.62	38.47 ± 1.14
CACTUs-MA-Bi [15]	57.84 ± 0.75	78.12 ± 0.67	34.98 ± 0.57	57.75 ± 0.58	36.13 ± 1.07
CACTUs-Pr-Bi [15]	53.58 ± 0.65	71.21 ± 0.68	32.79 ± 0.53	50.12 ± 0.51	36.05 ± 1.06
PsCo [16]	93.25 ± 0.59	97.56 ± 0.34	82.06 ± 0.43	91.01 ± 0.45	42.90 ± 0.95
Meta-GMVAE [23]	93.81 ± 0.75	96.85 ± 0.50	81.29 ± 0.62	89.00 ± 0.51	41.78 ± 1.13
MINO	93.75 ± 0.46	97.71 ± 0.37	83.57 ± 0.41	94.69 ± 0.40	44.73 ± 1.01
MAML (supervised) [13]	94.46	98.83	84.6	96.29	46.81