Boosting Transductive Few-Shot Fine-tuning with Margin-based Uncertainty Weighting and Probability Regularization

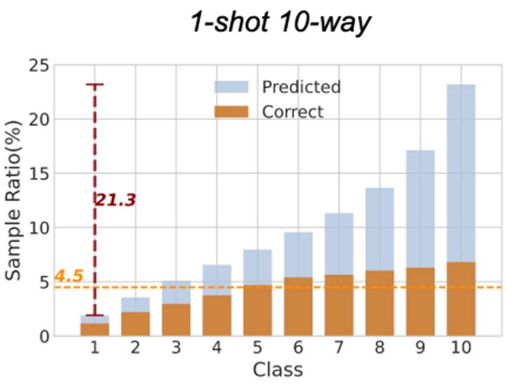
Ran Tao, Hao Chen, Marios Savvides
Carnegie Mellon University

WED-PM-324

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Method Overview

The Key Observation:

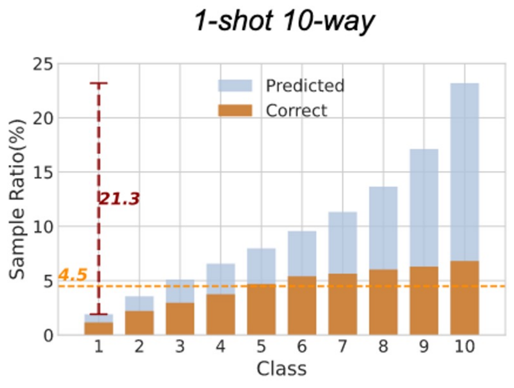


Even with the same number of per-class training samples, there is a severely imbalanced categorical performance:

1. Imbalanced #per-class predictions.
2. Imbalanced per-class accuracy.

Method Overview

The Key Observation:



Even with the same number of per-class training samples, there is a severely imbalanced categorical performance:

1. Imbalanced #per-class predictions.
2. Imbalanced per-class accuracy.

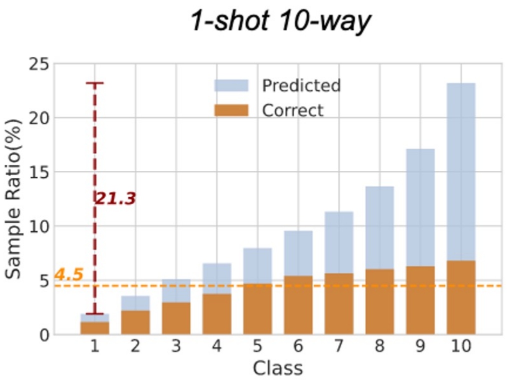
Transductive Fine-tuning[4]:

$$\theta^*(\mathcal{D}_s, \mathcal{D}_q) = \arg_\theta \min\left(\frac{1}{N_s} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_s} \mathcal{L}_s(\mathbf{x}, \mathbf{y}) + \frac{1}{N_q} \sum_{(\mathbf{x}) \in \mathcal{D}_q} \mathcal{L}_q(\mathbf{x})\right)$$

Add a loss for testing data

Method Overview

The Key Observation:



Without TF-MP, there is a severely imbalanced categorical performance even with the same number of per-class training samples.

Transductive Fine-tuning[4]:

$$\theta^*(\mathcal{D}_s, \mathcal{D}_q) = \arg_\theta \min\left(\frac{1}{N_s} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_s} \mathcal{L}_s(\mathbf{x}, \mathbf{y}) + \frac{1}{N_q} \sum_{(\mathbf{x}) \in \mathcal{D}_q} \mathcal{L}_q(\mathbf{x})\right)$$

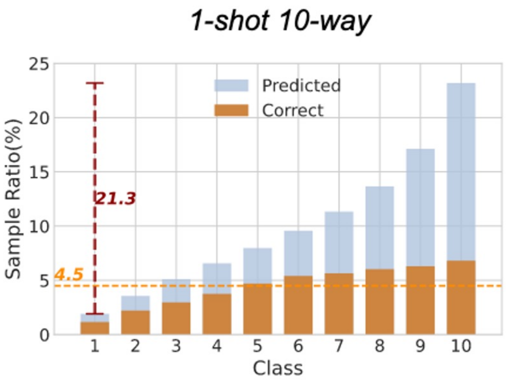
TF-MP:

Add a loss for testing data

$$\mathcal{L}_q(\mathbf{x}) = \lambda(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})) \times H(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x}))$$

Method Overview

The Key Observation:



Without TF-MP, there is a severely imbalanced categorical performance even with the same number of per-class training samples.

Transductive Fine-tuning[4]:

$$\theta^*(\mathcal{D}_s, \mathcal{D}_q) = \arg_\theta \min\left(\frac{1}{N_s}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_s} \mathcal{L}_s(\mathbf{x},\mathbf{y}) + \frac{1}{N_q}\sum_{(\mathbf{x})\in\mathcal{D}_q} \mathcal{L}_q(\mathbf{x})\right)$$

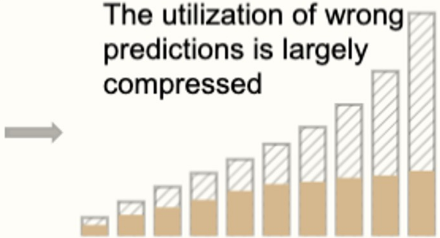
Add a loss for testing data

TF-MP:

$$\mathcal{L}_q(\mathbf{x}) = \boxed{\lambda(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x}))} \times H(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x}))$$

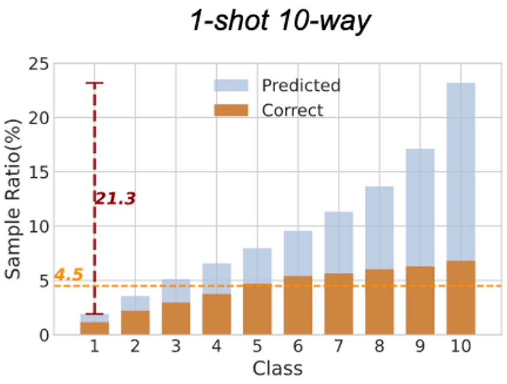
Margin-based Uncertainty Weighting

Weighting unlabeled testing data



Method Overview

The Key Observation:



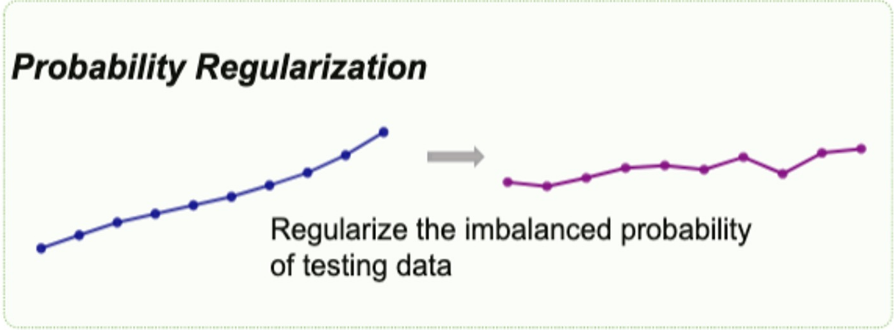
Without TF-MP, there is a severely imbalanced categorical performance even with the same number of per-class training samples.

Transductive Fine-tuning[4]:

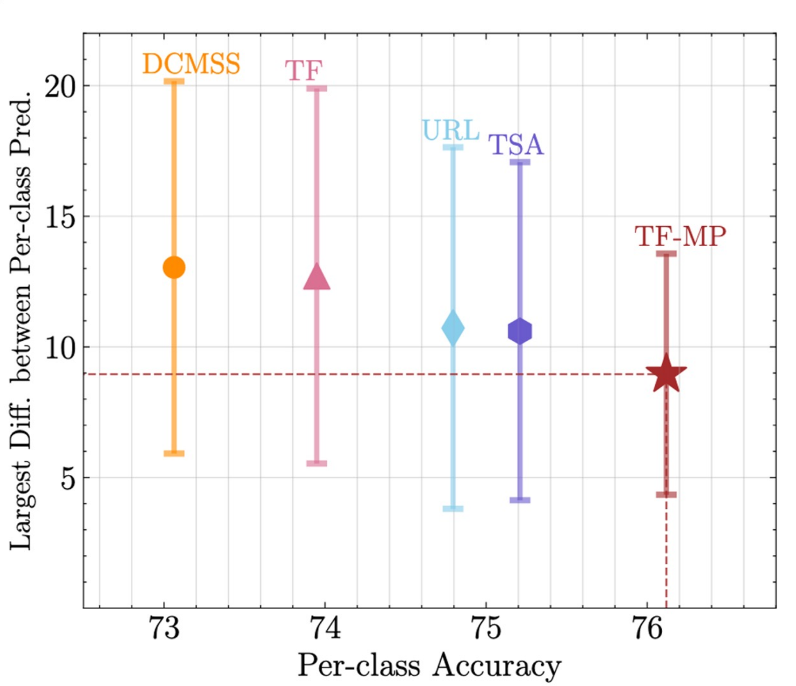
$$\theta^*(\mathcal{D}_s, \mathcal{D}_q) = \arg_\theta \min\left(\frac{1}{N_s} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_s} \mathcal{L}_s(\mathbf{x},\mathbf{y}) + \frac{1}{N_q} \sum_{(\mathbf{x})\in\mathcal{D}_q} \mathcal{L}_q(\mathbf{x})\right)$$

TF-MP:

$$\mathcal{L}_q(\mathbf{x}) = \lambda(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})) \times H(\boxed{\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})})$$



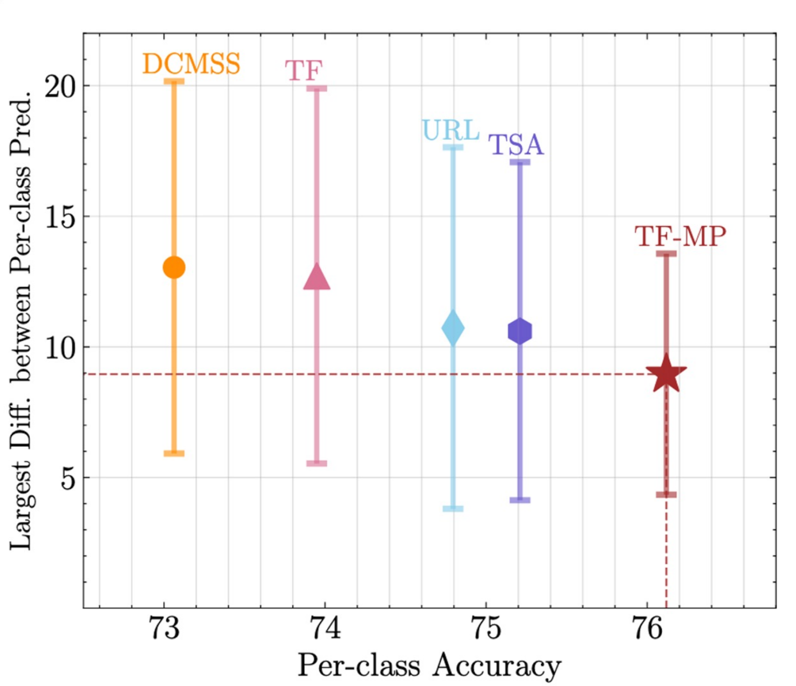
The Observation on Low-shot Fine-tuning



Performance of SOTA methods with a uniform testing set (10 per-class samples) using Meta-Dataset[1]:

- The Largest Difference (LD) between #per-class predictions is ideally 0 when each class is equally learned.
- LD is largely over 10 for SOTA methods.

The Observation on Low-shot Fine-tuning



This indicates: ***the learned class marginal distribution is largely imbalanced and biased.***

Solving this issue is critical to maintaining the algorithms' robustness in different testing scenarios.

Transductive Low-shot Fine-tuning

A few training samples

Fine-tuning

Involving testing data

Deep Model

Testing samples

Testing sample's prediction

Transductive Fine-tuning with Margin-based Uncertainty Weighting and Probability Regularization (TF-MP):

$$\mathcal{L}_q(\mathbf{x}) = \lambda(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})) \times H(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x}))$$

Margin-based uncertainty weighting

Probability Regularization

Margin-based Uncertainty Weighting

Given the predicted probability ***p*** for each unlabeled testing data, Entropy-based uncertainty is generally used to assign loss weights:

$$e(\mathbf{p}) = -\frac{\sum_i^C (p_i \log p_i)}{\log C}$$

Larger uncertainty refers to smaller loss weight:

$$\lambda(\mathbf{p}) = 1 - e(\mathbf{p})$$

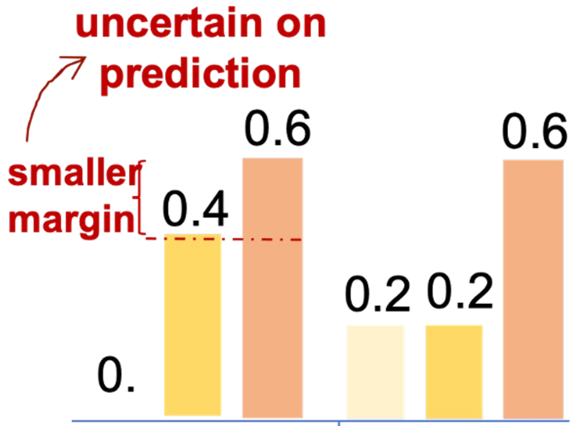
where $\mathbf{p}$ is the abbreviation for $\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})$

$\sum_i^c p_i = 1, \mathbf{p} = [p_1, p_2, ...p_c]$

C : the number of classes.

Margin-based Uncertainty Weighting

We emphasize the importance of ***the margin between the maximum and second maximum probability*** Δp in uncertainty computation.

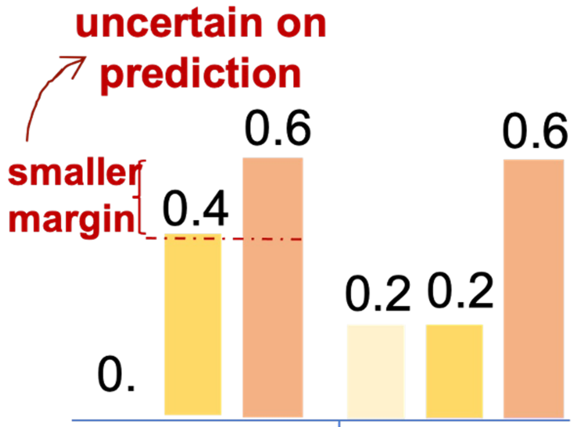


Confidence	0.6	0.6
Entropy	0.61	0.86

The entropy uncertainty cannot reflect the margin information.

Margin-based Uncertainty Weighting

We emphasize the importance of ***the margin between the maximum and second maximum probability*** Δp in uncertainty computation.

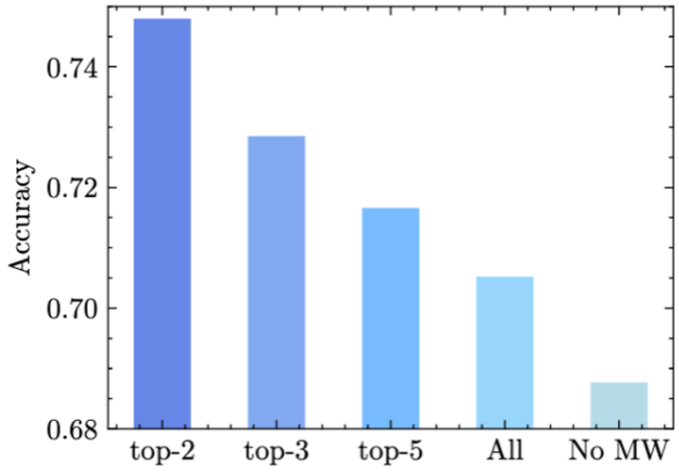


Confidence	0.6	0.6
Entropy	0.61	0.86
Margin-based Entropy	**0.97**	0.81

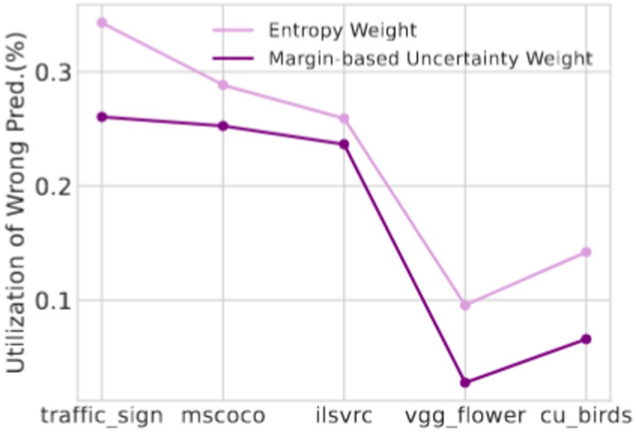
Margin-based Uncertainty:

$$\hat{e}(\mathbf{p}) = -\frac{1}{\log 2}(\hat{p}_{max}\log\hat{p}_{max} + (\hat{p}_{max} - \hat{\Delta p})\log(\hat{p}_{max} - \hat{\Delta p}))$$

Margin-based Uncertainty Weighting



(a) Weights with Top-k Prob.



- Margin-based Entropy (top-2) weighting outperforms Entropy weighting (All).
- The utilization of testing data with wrong predictions are largely compressed by Margin-based uncertainty weighting.

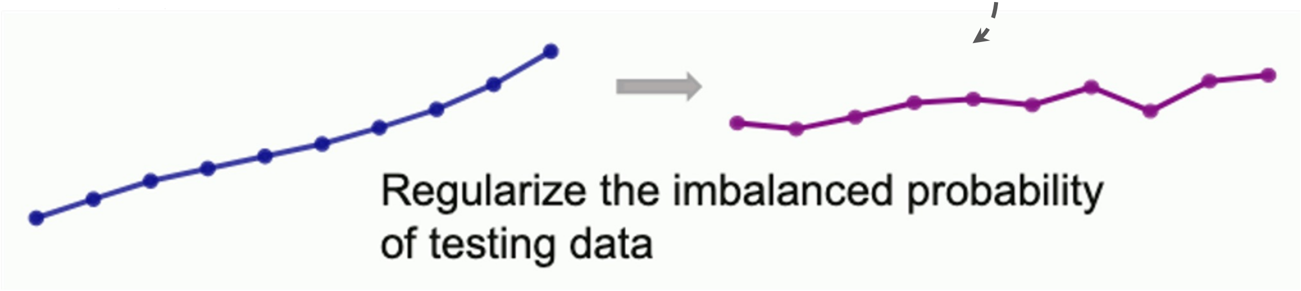
Probability Regularization

The loss objective for unlabeled testing data:

$$\mathcal{L}_q(\mathbf{x}) = \lambda(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})) \times H(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x}))$$

$$H(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x})) = -\hat{\mathbf{y}} \log(\mathbf{p}_\theta(\mathbf{y}|\mathbf{x}))$$

$\hat{\mathbf{y}}$: Pseudo-label for unsupervised testing data



Probability Regularization

The scale vector for each testing sample **x**:

$$\mathbf{v} = \frac{U}{\hat{E}_{x \cup \mathcal{D}_s}[p_\theta(\mathbf{y}|\mathbf{x})]}$$

$U \in \mathbb{R}^C$: uniform distribution

- The vector v quantifies the difference between uniform and the learned marginal distribution.
- The learned marginal is estimated using the set x∪D_s

Probability Regularization

The scale vector for each testing sample **x**:

$$\mathbf{v} = \frac{U}{\hat{E}_{x \cup \mathcal{D}_s}[p_\theta(\mathbf{y}|\mathbf{x})]}$$

$U \in \mathbb{R}^C$: uniform distribution

- For each testing data, the predicted probability q is regularized by element-wisely multiplied using v as follows:

$$\tilde{\mathbf{q}} = \text{Normalize}(\mathbf{q} * \mathbf{v})$$

$$\text{Normalize}(x_i) = \frac{x_i}{\sum_j x_j}$$

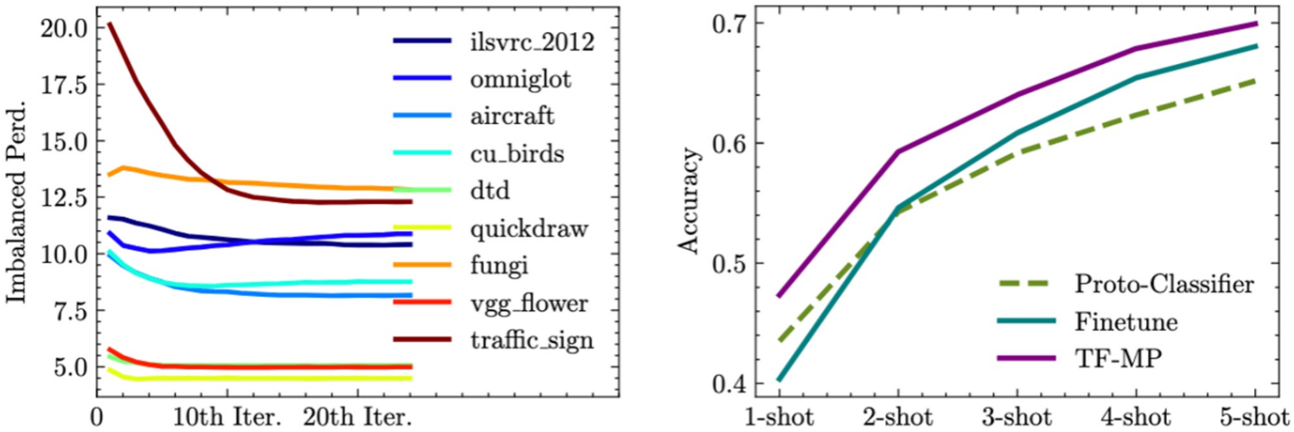
Experimental Results

We compare the state-of-the-art methods and benchmark on a published ***Meta-Dataset***[1]

Method	Backbone	ILSVRC	Omni	Acraft	Birds	DTD	QDraw	Fungi	Flower	Sign	COCO
fo-P-M [32]	–	49.5 ± 1.1	60.0 ± 1.4	53.1 ± 1.0	68.8 ± 1.0	66.6 ± 0.8	49.0 ± 1.1	39.7 ± 1.1	85.3 ± 0.8	47.1 ± 1.1	41.0 ± 1.1
BOHB [26]	-	51.9 ± 1.1	67.6 ± 1.2	54.1 ± 0.9	70.7 ± 0.9	68.3 ± 0.8	50.3 ± 1.0	41.4 ± 1.1	87.3 ± 0.6	51.8 ± 1.0	48.0 ± 1.0
LR [31]	ResNet18	60.1	64.9	63.1	77.7	78.6	62.5	47.1	91.6	77.5	57.0
Meta-B [6]	ResNet18	59.2	69.1	54.1	77.3	76.0	57.3	45.4	89.6	66.2	55.7
CNAPS [1]	ResNet18	54.8	62.0	49.2	66.5	71.6	56.6	37.5	82.1	63.1	45.8
DCM-S [30]	ResNet34	64.6	81.8	79.7	85.0	77.9	69.3	49.3	93.2	88.7	57.7
CTX [8]	ResNet34	62.7 ± 1.0	82.2 ± 1.0	79.5 ± 0.9	80.6 ± 0.9	75.6 ± 0.6	**72.7** ± 0.8	51.6 ± 1.1	**95.3** ± 0.4	82.6 ± 0.8	59.9 ± 1.0
TSA [17]	ResNet34	63.7 ± 1.0	82.6 ± 1.1	80.13 ± 1.0	83.4 ± 0.8	79.6 ± 0.7	71.0 ± 0.8	51.4 ± 1.2	94.1 ± 0.5	81.7 ± 1.0	61.7 ± 1.0
T-CNAPS [1]	ResNet18	54.1 ± 1.1	62.9 ± 1.3	48.4 ± 0.9	67.3 ± 0.9	72.5 ± 0.7	58.0 ± 1.0	37.7 ± 1.1	82.8 ± 0.8	61.8 ± 1.1	45.8 ± 1.0
T-F [7]	WRN-28	60.5	82.0	72.4	82.1	80.5	57.4	47.7	92.0	64.4	42.9
TF-MP	ResNet18	62.2 ± 1.1	83.8 ± 1.1	70.9 ± 0.9	81.3 ± 0.8	79.2 ± 0.6	70.5 ± 0.6	51.2 ± 1.0	93.3 ± 0.4	78.2 ± 1.0	**62.5** ± 0.9
TF-MP	ResNet34	**66.4** ± 1.0	**87.5** ± 0.8	**80.0** ± 0.9	**87.4** ± 0.6	**81.9** ± 0.6	71.9 ± 0.4	**54.9** ± 0.9	94.8± 0.4	**89.2** ± 0.9	61.5 ± 0.9

- TF-MP achieves SOTA performance over transductive settings in Meta-Dataset.
- TF-MP is effective with different scales of models and datasets from different domains

Experimental Results



(b) Imbalanced Pred. vs. Iteration.

(c) N-shot Analysis.

b)TF-MP effectively reduces the imbalance in per-class predictions during fine-tuning for various datasets.

c)TF-MP boosts performance over the different number of few-shot settings.

Thank you!

[1] Triantafillou, Eleni, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin et al. "Meta-dataset: A dataset of datasets for learning to learn from few examples."arXiv preprint arXiv:1903.03096(2019).

[2] Iscen, Ahmet, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. "Label propagation for deep semi-supervised learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5070-5079. 2019.

[3] Scheffer, Tobias, Christian Decomain, and Stefan Wrobel. "Active hidden markov models for information extraction." In International Symposium on Intelligent Data Analysis, pp. 309-318. Springer, Berlin, Heidelberg, 2001.

[4] Dhillon, Guneet S., Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. "A baseline for few-shot image classification." arXiv preprint arXiv:1909.02729 (2019).

[5] Tao, Ran, Han Zhang, Yutong Zheng, and Marios Savvides. "Powering Finetuning in Few-shot Learning: Domain-Agnostic Feature Adaptation with Rectified Class Prototypes." arXiv preprint arXiv:2204.03749 (2022).

[6] Li, Wei-Hong, Xialei Liu, and Hakan Bilen. "Cross-domain few-shot learning with task-specific adapters." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7161-7170. 2022.

[7] Li, Wei-Hong, Xialei Liu, and Hakan Bilen. "Universal representation learning from multiple domains for few-shot classification." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9526-9535. 2021.