Discriminative Sample-Guided and Parameter-Efficient Feature Space Adaptation for Cross-Domain Few-Shot Learning

Supplementary Material

9. Datasets

Meta-Dataset [55] is a few-shot classification benchmark that initially consists of ten datasets: ILSVRC_2012 (ImageNet) [51], Omniglot [34], FGVC-Aircraft (Aircraft) [44], CUB-200-2011 (Birds) [58], Describable Textures (Textures) [13], QuickDraw [31], FGVCx Fungi (Fungi) [8], VGG Flower [46], Traffic Signs [26] and MS-COCO [42]. It further expands with the addition of MNIST [35], CIFAR-10 [33] and CIFAR-100 [33]. Each dataset is further divided into train, validation and test sets with disjoint classes. We follow the standard training protocols proposed by [55] and consider both "Training on all datasets" (MDL: multi-domain learning) and "Training on ImageNet-Train only" (SDL: single-domain learning) settings. For the former, we follow the standard procedure and use the training set of the first eight datasets for pre-training. During evaluation, the test set of the eight datasets are used for evaluating the generalization ability in the seen domains while the remaining five datasets are used to evaluate the cross-domain generalization ability. In the "Training on ImageNet-Train only" setting, we follow the standard procedure and only use the train set of ImageNet for pre-training. The evaluation is performed on the test set of ImageNet as the seen domain while the rest 12 datasets are considered unseen domains. Additionally, to compare our method with more recent state-of-the-art [28], we also use a pre-trained model on the full ImageNet dataset for "Training on ImageNet-Full" (SDL-E: single-domain learning-extra data) setting, where the evaluation is performed similarly to the "Training on ImageNet-Train only" setting.

We also report additional results for the following datasets in the Appendix.

miniImageNet [57] contains 100 classes from ImageNet-1k, set into 64 training, 16 validation and 20 testing classes.

CIFAR-FS [5] is created by dividing the original CIFAR-100 into 64 training, 16 validation and 20 testing classes.

10. Implementation details

10.1. Pre-trianing using Masked Image Modelling

We employ Masked Image Modelling (MIM) to pre-train the feature extractor and follow the hyper-parameters and data augmentations recommended in [62]. The teacher patch temperature was set to 0.04, in contrast to the default value of 0.07 after observing that a lower temperature leads to more consistent and stable training losses.[2].

MDL: We employ the train sets of the eight in-domain datasets (ImageNet, Omniglot, Aircraft, Birds, Textures, and VGG Flower) considered under the MDL setting for pre-training f_θ.

SDL: We employ the train set of the ImageNet dataset for pre-training f_θ.

SDL-E: The entire ImageNet dataset is utilized to train the feature extractor [3] f_θ [28].

10.2. Pre-trianing on DINO

SDL-E: To compare with the MIM pre-training on SDL-E, we utilize the pre-trained checkpoint weights provided by DINO [9] after training on the entire ImageNet dataset.

10.3. Hyperparameters

10.3.1 Task-specific parameter initialization

For results reported in the main text, we choose constant initialization of task-specific parameters γ (scale) and β (shift) as one and zero, respectively. However, one could also employ a normalized initialization, where the mean values of γ and β are one and zero [62]. Therefore, we report results for normalized initialization and constant initialization in Supplementary Table 6 columns 1 and 2, respectively. Notably, we obtain better results for the constant initialization in comparison to the normalized initialization.

10.3.2 AdamW vs NAdam

Recently, AdamW has gained popularity as a preferred choice for fine-tuning large models such as ViTs [59]. Nevertheless, our results, as detailed in Supplementary Table 6, columns 2 and 3, demonstrate that NAdam yields superior performance in the context of cross-domain few-shot classification.

10.3.3 Anchor initialization

While A_ϕ anchors in DIPA are randomly initialized for each task, one can argue that using the mean of the support embedding vectors can be a favourable anchor initialization point. Consequently, we report the results for random vs custom initialization of anchors in Supplementary Table 6, columns 2 and 4. Here, the anchors are randomly initialized for A_ϕ (random) while the mean of class embedding vectors

[2] https://github.com/bytedance/ibot/issues/19

[3] https://github.com/bytedance/ibot#pre-trained-models

γ, β (constant)		✓	✓	✓
γ, β (normal)	✓			
NAdam	✓	✓		✓
AdamW			✓	
A_ϕ (random)	✓	✓	✓	
A_ϕ (custom)				✓
ImageNet	69.22 ± 0.94	**70.86 ± 0.95**	68.21 ± 0.96	70.25 ± 0.98
Omniglot	83.55 ± 1.17	**84.68 ± 1.10**	82.79 ± 1.18	84.55 ± 1.15
Aircraft	85.91 ± 1.06	86.33 ± 0.95	**86.55 ± 1.00**	85.05 ± 1.06
Birds	90.31 ± 0.80	**90.75 ± 0.75**	88.49 ± 0.88	89.70 ± 0.88
Textures	87.66 ± 0.66	88.60 ± 0.51	87.15 ± 0.62	**88.61 ± 0.56**
Quick Draw	74.27 ± 0.82	**75.29 ± 0.77**	72.81 ± 0.83	75.10 ± 0.77
Fungi	66.07 ± 1.05	**66.64 ± 1.05**	64.30 ± 1.05	65.54 ± 1.07
VGG Flower	97.71 ± 0.32	**97.88 ± 0.30**	97.24 ± 0.38	97.63 ± 0.32
Traffic Sign	89.84 ± 1.19	**91.29 ± 0.96**	87.29 ± 1.13	89.80 ± 0.97
MS-COCO	62.34 ± 1.04	**64.78 ± 0.95**	57.84 ± 1.07	64.67 ± 1.01
MNIST	96.64 ± 0.49	**96.87 ± 0.53**	96.14 ± 0.60	96.82 ± 0.50
CIFAR-10	84.56 ± 0.85	87.40 ± 0.64	79.72 ± 1.03	**87.81 ± 0.66**
CIFAR-100	79.38 ± 0.94	**81.24 ± 0.78**	75.29 ± 0.94	80.28 ± 0.83
Average Seen	79.6	**82.6**	80.9	82.1
Average Unseen	82.6	**84.3**	79.3	82.7
Average All	80.8	**83.3**	80.3	82.3

Table 6. Comparision of varying the task-specific parameter initialization (constant vs normal), Optimizers (NAdam vs AdamW) and A_ϕ anchor initialization (random vs custom) in the MDL setting.

initializes the anchors for A_ϕ (custom). Notably, using unadapted feature embeddings for anchor initialization may hinder fine-tuning due to priors imposed by unadapted features. In contrast, using random initialization, together with a substantial learning rate may offer better adaptability for anchors during fine-tuning without being influenced by irrelevant priors. This is also reflected in the results reported in Supplementary Table 6, columns 2 and 4, where random initialization outperforms custom initialization, confirming our selection in the DIPA framework.

10.3.4 Number of fine-tuning iterations

We experimentally determined the number of fine-tuning iterations. We report the results for four such scenarios in Supplementary Table 7. As reported in the results, l_{A_ϕ} with 80 iterations provides the highest accuracy. Therefore, in our framework, we use 80 as the number of iterations for fine-tuning.

Fine Tune	# Iterations	Avg. Seen	Avg. Unseen	Avg. All
DIPA	40	81.5	82.4	81.8
DIPA	**80**	**82.6**	**84.3**	**83.3**

Table 7. Comparing the average (Avg.) performance variation as the number of epochs varies for seen, unseen and all domains in the MDL setting.

11. Addtional Results for Meta-Dataset

11.1. Feature Space Visualizations: Before and After Fine-Tuning

By using UMAP visualizations, we identify the impact of fine-tuning the feature space using l_{A_ϕ} in Supplementary Fig. 6 and 7. Here, the left columns illustrate that semantic clusters have already emerged using the pre-trained MIM features, although overlapped/dispersed in some instances. Thereafter, as illustrated in the right columns, l_{A_ϕ} uses the strong initialization provided by MIM and further refines the feature space to form better-separated clusters that show high inter-class variance and low intra-class variance.

11.2. Prototype Visualizations

The placement of anchors A_ϕ and mean embedding-based prototypes after fine-tuning is visualized in Fig. 8. As discussed in Section 6.1, while A_ϕ provides strong supervision for cluster formation during fine-tuning, after fine-tuning, we observe that they are placed with a small offset from the mean representation (mean embedding prototype) of the clusters.

11.3. Impact of tuned depth

Supplementary Table 8 reports the variation of accuracies as the number of tuned layers d_t vary on the MDL setting for Meta-Dataset. A summary of Supplementary Table 8 is shown in the main text's Fig. 4 and 5.

11.4. Feature fusion depth

We report the dataset-level accuracy values obtained as we vary the feature fusion depths in Supplementary Table 9, where a summary of it was presented in the main text.

11.5. Pre-training results

The dataset-level accuracies reported in the SDL-E setting by DINO and MIM pre-trained models with varying fine-tuning strategies are reported in Supplementary Table 10.

11.6. Further results on Meta-Dataset

After evaluating our framework over a broad range of varying shots K (e.g. up to 100 shots), we further analyze our framework in a more challenging setting. While l_{A_ϕ} requires at least two examples per class in order to gain benefits from its discriminative sample-based feature space adaptation, here we evaluate its performance in the more challenging varying-way, 5-shot setting, comparing it with other works that have reported results in this context [37]. As shown in Supplementary Table 11, overall performance for all methods has decreased due to the even more challenging nature of the support set. Nevertheless, our method still outperforms the existing methods when the number of

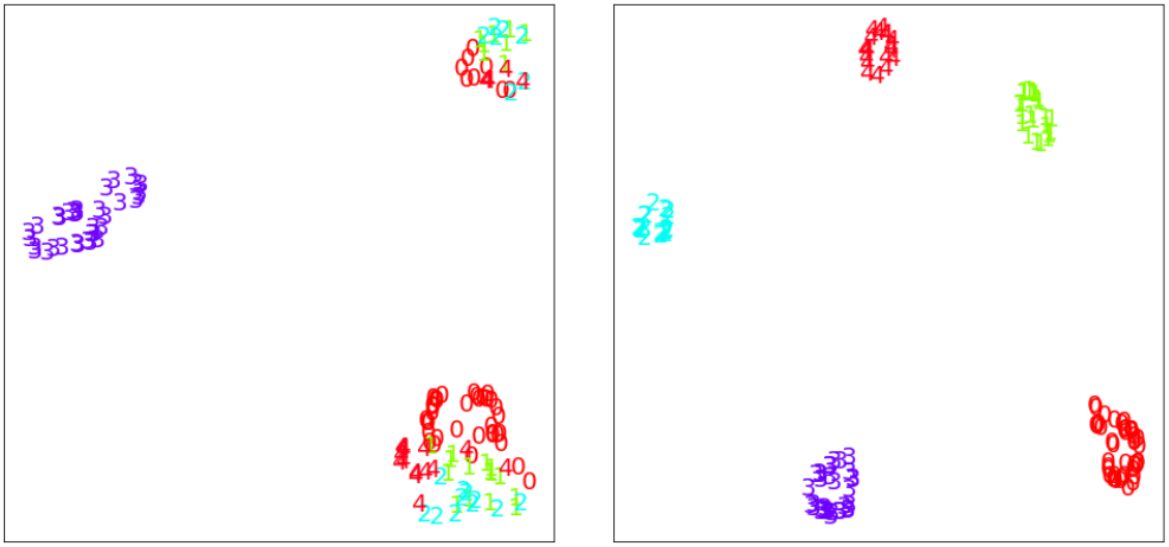

Figure 6. UMAP visualization of clusters formed in the feature space for Aircraft domain in MDL setting. The clusters formed before and after fine-tuning with DIPA, are illustrated in the first and second columns, respectively.

d	0	1	2	3	4	5	6	7	8	9	10	11	12
ImageNet	66.51 ± 1.02	68.22 ± 0.95	70.24 ± 1.01	71.11 ± 0.95	71.37 ± 0.94	71.00 ± 0.92	71.36 ± 0.91	70.86 ± 0.95	69.71 ± 0.94	68.39 ± 0.95	68.05 ± 0.92	67.57 ± 0.95	67.13 ± 0.93
Omniglot	67.04 ± 1.23	72.63 ± 1.29	71.87 ± 1.34	75.52 ± 1.25	80.10 ± 1.16	81.92 ± 1.19	83.58 ± 1.09	84.68 ± 1.10	82.91 ± 1.25	84.25 ± 1.19	84.51 ± 1.16	84.81 ± 1.11	84.33 ± 1.16
Aircraft	52.97 ± 0.95	75.97 ± 0.96	77.01 ± 1.04	80.67 ± 0.99	83.88 ± 0.99	85.12 ± 0.99	85.95 ± 1.02	86.33 ± 0.95	85.45 ± 1.09	86.35 ± 0.95	85.04 ± 1.03	85.35 ± 0.99	83.44 ± 1.12
Birds	83.12 ± 0.82	89.04 ± 0.69	90.40 ± 0.59	90.20 ± 0.67	90.92 ± 0.71	91.01 ± 0.74	91.22 ± 0.67	90.75 ± 0.75	90.50 ± 0.68	89.74 ± 0.77	89.17 ± 0.80	89.56 ± 0.74	88.63 ± 0.75
Textures	84.95 ± 0.50	86.89 ± 0.58	88.34 ± 0.53	88.49 ± 0.52	88.25 ± 0.59	88.95 ± 0.56	88.52 ± 0.58	88.60 ± 0.51	87.83 ± 0.65	87.39 ± 0.64	86.20 ± 0.66	84.95 ± 0.71	85.47 ± 0.57
Quickdraw	54.78 ± 0.94	63.47 ± 0.93	64.63 ± 0.96	67.80 ± 0.90	71.55 ± 0.94	73.69 ± 0.85	74.65 ± 0.85	75.29 ± 0.77	74.38 ± 0.75	75.55 ± 0.80	75.97 ± 0.73	75.49 ± 0.75	75.05 ± 0.83
Fungi	57.33 ± 1.06	61.30 ± 1.06	62.99 ± 1.18	63.86 ± 1.07	66.35 ± 1.08	66.87 ± 1.12	66.91 ± 1.06	66.64 ± 1.05	67.52 ± 1.05	65.03 ± 1.06	63.45 ± 1.10	63.59 ± 1.06	63.96 ± 1.08
VGG_Flower	97.56 ± 0.25	97.10 ± 0.34	97.29 ± 0.33	97.27 ± 0.36	97.78 ± 0.30	98.06 ± 0.28	97.99 ± 0.29	97.88 ± 0.30	97.65 ± 0.31	97.55 ± 0.32	97.14 ± 0.36	97.00 ± 0.38	97.28 ± 0.32
Traffic Sign	40.20 ± 1.07	52.42 ± 1.25	57.42 ± 1.27	60.77 ± 1.19	68.90 ± 1.23	75.15 ± 1.17	80.71 ± 1.17	85.52 ± 1.04	89.46 ± 0.91	91.29 ± 0.96	91.99 ± 0.96	92.76 ± 0.83	92.25 ± 0.96
MSCOCO	54.13 ± 0.97	56.50 ± 0.95	58.64 ± 0.97	63.19 ± 1.01	63.22 ± 0.95	65.75 ± 1.00	65.59 ± 0.97	65.32 ± 0.93	64.70 ± 0.97	64.78 ± 0.95	63.46 ± 0.92	62.56 ± 1.00	62.02 ± 0.94
MNIST	74.81 ± 0.74	86.77 ± 0.75	88.46 ± 0.73	89.86 ± 0.80	93.75 ± 0.68	94.51 ± 0.68	95.27 ± 0.65	96.13 ± 0.59	96.68 ± 0.53	96.87 ± 0.53	96.78 ± 0.50	97.51 ± 0.40	97.12 ± 0.45
CIFAR-10	81.54 ± 0.64	86.54 ± 0.60	87.14 ± 0.65	87.60 ± 0.59	88.39 ± 0.58	89.17 ± 0.60	88.92 ± 0.61	89.04 ± 0.56	88.09 ± 0.61	87.40 ± 0.64	86.43 ± 0.68	85.75 ± 0.71	84.70 ± 0.76
CIFAR-100	73.41 ± 0.88	78.02 ± 0.79	78.57 ± 0.75	79.08 ± 0.78	81.32 ± 0.76	80.71 ± 0.76	80.99 ± 0.78	81.33 ± 0.81	81.45 ± 0.71	81.24 ± 0.78	80.02 ± 0.77	78.78 ± 0.79	78.53 ± 0.81
Average Seen	70.5	76.8	77.8	79.4	81.3	82.1	82.5	**82.6**	82	81.8	81.2	81	80.7
Average Unseen	64.8	72	74	76.1	79.1	81.1	82.3	83.5	84.1	**84.3**	83.7	83.5	82.9
Average All	68.3	75	76.4	78.1	80.4	81.7	82.4	**83**	82.8	82.8	82.2	82	81.5

Table 8. Variation of accuracies as the number of tuned layers d_t varies in the MDL setting for in-domain and out-of-domain datasets in Meta-Dataset.

Fusion depth	ImageNet	Omniglot	Aircraft	Birds	Textures	Quick Draw	Fungi	VGG Flower	Traffic Sign	MS-COCO	MNIST	CIFAR-10	CIFAR-100	Average All
1	70.2 ± 0.9	84.9 ± 1.1	86.2 ± 1.1	90.5 ± 0.7	88.3 ± 0.6	74.7 ± 0.8	66.5 ± 1.0	97.3 ± 0.4	90.3 ± 1.0	63.1 ± 0.9	**97.4 ± 0.4**	87.6 ± 0.6	80.4 ± 0.8	82.9
2	70.5 ± 1.0	84.4 ± 1.2	86.7 ± 1.0	90.8 ± 0.7	87.9 ± 0.6	74.8 ± 0.8	66.6 ± 1.1	97.5 ± 0.3	90.4 ± 1.0	63.3 ± 1.0	97.0 ± 0.5	**88.0 ± 0.6**	80.0 ± 0.8	82.9
4	70.9 ± 1.0	84.7 ± 1.1	**86.3 ± 1.0**	90.8 ± 0.8	**88.6 ± 0.5**	**75.3 ± 0.8**	**66.6 ± 1.0**	**97.9 ± 0.3**	**91.3 ± 1.0**	**64.8 ± 1.0**	96.9 ± 0.5	87.4 ± 0.6	**81.2 ± 0.8**	**83..3**
6	**71.6 ± 0.9**	**85.5 ± 1.1**	86.1 ± 1.0	**90.9 ± 0.7**	88.0 ± 0.5	**75.3 ± 0.8**	66.0 ± 1.0	97.6 ± 0.4	90.3 ± 1.1	63.7 ± 1.0	97.3 ± 0.4	87.3 ± 0.6	80.8 ± 0.8	83.1
8	69.2 ± 1.0	85.4 ± 1.1	85.6 ± 1.0	90.4 ± 0.8	87.4 ± 0.6	75.0 ± 0.8	65.4 ± 1.2	97.6 ± 0.4	90.7 ± 1.0	63.5 ± 1.0	96.9 ± 0.6	86.3 ± 0.7	80.0 ± 0.8	82.6
12	68.0 ± 1.0	85.4 ± 1.1	85.7 ± 0.9	89.4 ± 0.9	87.7 ± 0.6	74.5 ± 0.8	63.6 ± 1.2	97.7 ± 0.3	**91.3 ± 0.9**	62.7 ± 1.0	97.2 ± 0.4	86.1 ± 0.6	78.6 ± 0.8	82.1

Table 9. Variation of accuracies as the feature fusion depth d_f vary on the MDL setting.

support images per class is fewer, especially on the challenging unseen domains by 2.9%.

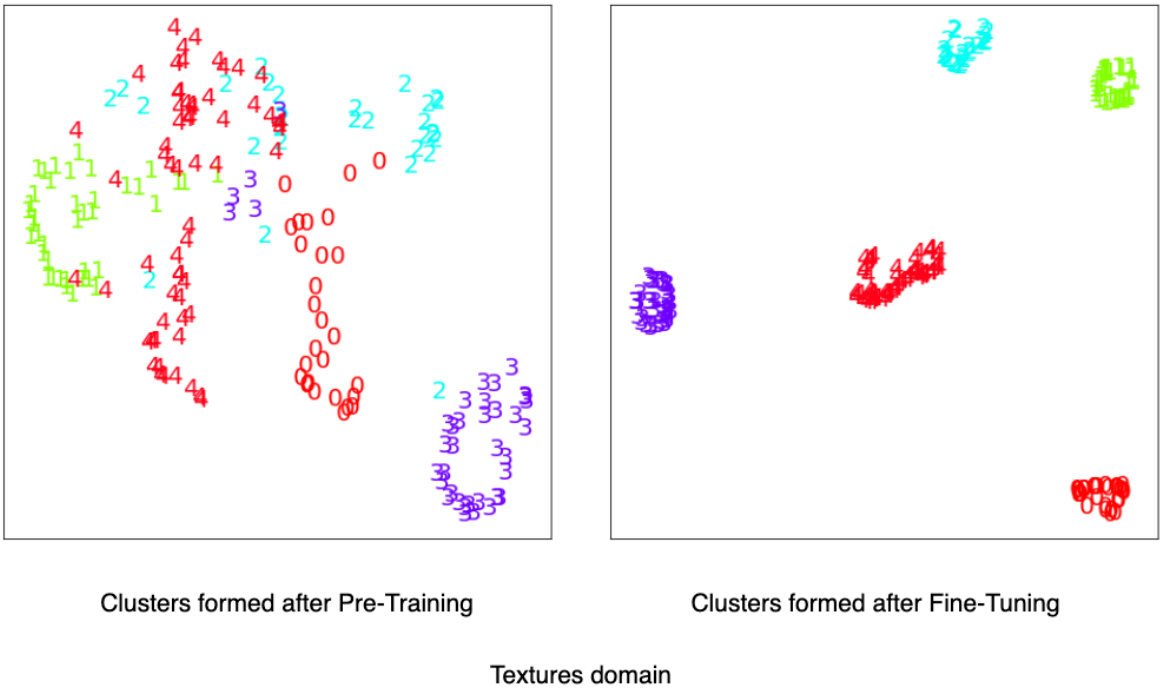

Figure 7. UMAP visualization of clusters formed in the feature space for Textures domain in MDL setting. The clusters formed before and after fine-tuning with DIPA are illustrated in the first and second columns, respectively.

Pre-training	MIM		DINO	
Fine-tuning	NCC	DIPA	NCC	DIPA
ImageNet	75.71 ± 0.81	77.26 ± 0.74	75.47 ± 0.82	75.89 ± 0.78
Omniglot	80.38 ± 1.36	84.06 ± 1.20	80.19 ± 1.31	83.65 ± 1.15
Aircraft	83.06 ± 1.03	87.09 ± 0.99	81.41 ± 1.10	85.88 ± 1.00
Birds	88.32 ± 0.75	90.52 ± 0.67	87.91 ± 0.77	90.37 ± 0.65
Textures	86.23 ± 0.69	87.32 ± 0.63	86.51 ± 0.72	87.06 ± 0.63
Quickdraw	73.38 ± 0.81	75.41 ± 0.81	72.62 ± 0.88	75.30 ± 0.79
Fungi	59.57 ± 1.08	60.89 ± 1.09	60.02 ± 1.14	62.16 ± 1.10
VGG_Flower	96.88 ± 0.40	97.48 ± 0.36	96.56 ± 0.41	97.24 ± 0.35
Traffic Sign	89.93 ± 0.94	91.66 ± 0.84	89.68 ± 0.94	91.20 ± 0.81
MSCOCO	64.52 ± 0.98	66.54 ± 0.93	64.30 ± 0.96	65.13 ± 0.99
MNIST	96.15 ± 0.50	97.24 ± 0.45	95.21 ± 0.63	96.82 ± 0.49
CIFAR-10	90.23 ± 0.66	92.23 ± 0.47	88.20 ± 0.76	89.95 ± 0.66
CIFAR-100	82.21 ± 0.79	84.48 ± 0.70	80.97 ± 0.76	82.29 ± 0.76
Average Seen	75.7	**77.3**	75.5	75.9
Average Unseen	82.6	**84.6**	82.0	83.9
Average All	82.0	**84.0**	81.5	83.3

Table 10. The impact of varying the pre-training and finetuning algorithms in SDL-E setting.

	Simple CNAPS	SUR	URT	TSA	DIPA
SS PT					✓
Sup. MT	✓	✓	✓	✓	
Backbone	RN18	RN18	RN18	RN18	ViT-s
ImageNet	47.2 ± 1.0	46.7 ± 1.0	48.6 ± 1.0	48.3 ± 1.0	**60.17 ± 0.80**
Omniglot	95.1 ± 0.3	95.8 ± 0.3	96.0 ± 0.3	**96.8 ± 0.3**	91.30 ± 0.46
Aircraft	74.6 ± 0.6	82.1 ± 0.6	81.2 ± 0.6	**85.5 ± 0.5**	64.77 ± 0.68
Birds	69.6 ± 0.7	62.8 ± 0.9	71.2 ± 0.7	76.6 ± 0.6	**87.55 ± 0.39**
Textures	57.5 ± 0.7	60.2 ± 0.7	65.2 ± 0.7	68.3 ± 0.7	**79.69 ± 0.50**
Quickdraw	70.9 ± 0.6	79.0 ± 0.5	**79.2 ± 0.5**	77.9 ± 0.6	68.40 ± 0.68
Fungi	50.3 ± 1.0	66.5 ± 0.8	66.9 ± 0.9	**70.4 ± 0.8**	66.57 ± 0.77
VGG_Flower	86.5 ± 0.4	76.9 ± 0.6	82.4 ± 0.5	89.5 ± 0.4	**96.96 ± 0.18**
Traffic Sign	55.2 ± 0.8	44.9 ± 0.9	45.1 ± 0.9	72.3 ± 0.6	**83.91 ± 0.45**
MSCOCO	49.2 ± 0.8	48.1 ± 0.9	52.3 ± 0.9	56.0 ± 0.8	**64.64 ± 0.68**
MNIST	88.9 ± 0.4	90.1 ± 0.4	86.5 ± 0.5	**92.5 ± 0.4**	92.07 ± 0.33
CIFAR-10	66.1 ± 0.7	50.3 ± 1.0	61.4 ± 0.7	72.0 ± 0.7	**80.37 ± 0.53**
CIFAR-100	53.8 ± 0.9	46.4 ± 0.9	52.5 ± 0.9	64.1 ± 0.8	**76.79 ± 0.64**
Average Seen	69.0	71.2	73.8	76.7	**76.9** (+0.2)
Average Unseen	62.6	56.0	59.6	71.4	**74.3** (+2.9)
Average All	66.5	65.4	68.3	74.6	**76.4** (+2.2)

Table 11. Results of Varying-Way Five-Shot in the MDL setting. Average (Avg.) accuracies are reported. RN: ResNet, ViT-s: ViT-small, SS PT: indicates self-supervised pre-training and Sup. MT: indicates supervised meta-training.

12. Addtional Results for miniImageNet and CIFAR-FS

Supplementary Table 12 reports the results for evaluating the DIPA framework under the SDL-E setting on CIFAR-FS and mini-ImageNet datasets. Here, we follow PMF [28] and compare DIPA with relevant existing methods. Our approach can be directly compared with methods that employ SSL for pre-training, both with and without subse-

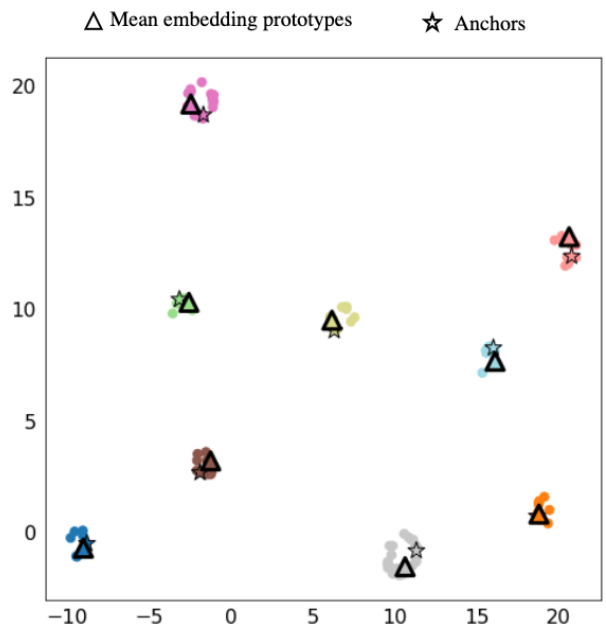

Figure 8. UMAP visualization of clusters formed in the feature space after fine-tuning with mean embedding-based prototypes and anchors A_ϕ.

quent fine-tuning (Supplementary Table 12, row D0-D4). Among those methods, our approach has superior performance across most scenarios. Notably, among the other methods that use various other training strategies, we still obtain somewhat good performance without requiring additional meta-training or training labels.

ID	Method	Backbone	Ext. dat.	Ext. lab.	miniImageNet 5/1	miniImageNet 5/5	CIFAR-FS 5/1	CIFAR-FS 5/5
	Inductive							
A0	Baseline++ [11]	CNN-4-64			48.2 ± 0.8	66.4 ± 0.6		
A1	MetaOpt-SVM [36]	ResNet12			62.6 ± 0.6	78.6 ± 0.5	72.0 ± 0.7	84.2 ± 0.5
A2	Meta-Baseline [12]	ResNet12			63.2 ± 0.2	79.3 ± 0.2		
A3	RS-FSL [1]	ResNet12		✓	65.3 ± 0.8			
	Transductive							
B0	Fine-tuning [15]	WRN-28-10			65.7 ± 0.7	78.4 ± 0.5	76.6 ± 0.7	85.8 ± 0.50
B1	SIB [27]	WRN-28-10			70.0 ± 0.6	79.2 ± 0.4	80.0 ± 0.6	85.3 ± 0.4
B2	PT-MAP [29]	WRN-28-10			82.9 ± 0.3	88.8 ± 0.1	**87.7 ± 0.2•**	90.7 ± 0.2
B3	CNAPS + FETI [4]	WRN-28-10	✓	✓	79.9 ± 0.8	91.5 ± 0.4		
	Semi-Supervised							
C0	LST [39]	ResNet12	✓		70.1 ± 1.9	78.7 ± 0.8		
C1	PLCM [30]	ResNet12	✓		72.1 ± 1.1	83.7 ± 0.6	77.6 ± 1.2	86.1 ± 0.7
	Self-Supervised							
D0	ProtoNet [23]	WRN-28-10			62.9 ± 0.5	79.9 ± 0.3	73.6 ± 0.3	86.1 ± 0.2
D1	ProtoNet [10]	AMDIM ResNet	✓		76.8 ± 0.2	91.0 ± 0.1		
D2	EPNet + SSL [50]	WRN-28-10	✓		79.2 ± 0.9	88.1 ± 0.5		
D3	FewTure [25]	ViT-small			68.0 ± 0.9	84.5 ± 0.5	**76.1 ± 0.9***	86.1 ± 0.6
D4	DIPA	ViT-small	✓		**79.6 ± 0.7***	**94.3 ± 0.3***	65.2 ± 0.9	**88.4 ± 0.6***
	Self-Supervised + MT							
E0	PMF [28]	ViT-small	✓		**93.1•**	**98.0•**	81.1	**92.5•**

Table 12. Comparison with representative state-of-the-art FSL algorithms on miniImageNet & CIFAR-FS for 5-way-1-shot (5/1) and 5-way-5-shot (5/5). Mean accuracy and 95% confidence interval are reported, where available. ✓ indicates the use of Extra data or Extra labels. MT: Meta-train, and * denotes the highest performance among the most relevant methods that are directly comparable to DIPA while • denotes the highest performance overall.