

Supplementary Experiments for CLIP-Adapter and PLOT

Anonymous Authors

1 SUPPLEMENTARY EXPERIMENTS

1.1 The difference between Ours and PLOT

We will briefly describe the difference between our method and PLOT.

Different Designs: In contrast to PLOT, we obtain a global text feature for **textual one prompt**, instead of a local text feature with **textual multi-prompt input**. The multi-prompt is that the set of learnable prompt is 4 ($N = 4$) and the set of output text feature is 4. The model optimizes four sentences simultaneously ("a photo of a dog", "a picture of a dog", "a drawing of a dog", "a good drawing of a dog"). Each embedding in each sentence participates in learning. The output text feature of our method is global feature ($N = 1$), the input sentence is **one text prompt** such as "a photo of a", and the text prompt is **non-learnable**. Our method is more simpler than **text multi-prompt input method (PLOT)** in textual design component. Further, the core component of PLOT is **OT**, while the core component of our method is **EnLa**. Our approach focuses on the proposal of EnLa and frozen text prompt, which is different with the design proposal of PLOT.

Different problems: Our approach primarily addresses **generalization issues**, such as base-to-novel and cross-data transfer experiments. In contrast, PLOT mainly focuses on **few-shot learning experiments** that tackle supervised tasks. However, our method also excels in few-shot learning experiments. As a result, the problems we aim to solve using optimal transport are **entirely distinct** from those addressed by PLOT.

1.2 End-to-end OT

The transport plan is efficiently computed through a limited number of matrix multiplications as a forward module. These matrix multiplications are crucial for determining the gradients that are then preserved for back-propagation. While the optimization strategy involves a two-stage process with optimal transport and prompts, the overall training flow remains **end-to-end**.

1.3 Analysis of parameters

In terms of the number of parameters in the overall model, our method has more parameters than PLOT because our method has EnLa and learnable visual embeddings. We have added experiments on CLIP-Adapter and PLOT.

1.4 Few-shot learning experiments for ViT-B/16 backbones

Our method provides improvements on few-shot settings compared to PLOT. This indicates that our method is more excellent.

1.5 Cross Dataset Evaluation for ViT-B/16 backbones

To verify the cross-dataset generalization ability, we train our method on the ImageNet dataset with 1,000 classes, and test it

Table 1: Few-shot learning experiments for ViT-B/16 backbones.

Dataset	PLOT	Ours
Average	82.09	83.31 ($\Delta +1.22$)
ImageNet	72.60	73.35
Caltech101	96.04	96.37
OxfordPets	93.59	94.03
StanfordCars	84.55	84.17
Flowers102	97.56	98.2
Food101	87.11	87.65
FGVCAircraft	46.74	50.91
SUN397	76.03	77.20
DTD	71.43	74.57
EuroSAT	92.00	93.13
UCF101	85.34	86.87

Table 2: Cross-dataset benchmark evaluation. Our method achieves overall favorable performance.

	Source		Target									
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
Linear probe CLIP	66.73	92.94	89.07	65.29	71.30	86.11	24.87	62.62	44.56	47.69	66.77	65.12
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CLIP-Adapter	71.40	93.85	89.57	64.66	68.85	85.54	18.53	64.35	41.86	46.43	66.77	64.04
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
PLOT	70.15	94.60	90.23	65.41	71.97	86.32	22.87	67.22	44.99	46.57	68.32	65.85
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
Ours	71.03	93.93	91.20	65.63	71.73	86.40	25.13	67.67	46.47	48.96	69.73	66.69

on the remaining 10 datasets. We have added experiments on CLIP-Adapter and PLOT. As shown in Table 2, our method shows competitive performance and achieves better generalization in 8/10 over the PLOT.

1.6 Domain Generalization Experiments for ViT-B/16 backbones

We have added experiments on CLIP-Adapter and PLOT. Compared with PLOT, our method shows improved performance in all ImageNet variants datasets.

Table 3: Domain generalization. These approaches are trained on imageNet and tested on datasets with domain shifts.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
Linear probe CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	71.51	64.2	47.99	49.71	75.21	59.28
CLIP-Adapter	71.40	64.5	47.72	49.75	75.55	59.38
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
PLOT	70.15	64.17	49.15	50.83	76.5	60.16
MaPLe	70.72	64.07	49.15	50.9	76.98	60.27
Ours	71.03	64.3	49.5	51.45	77.83	60.77

Table 5: Base-to-novel generalization experiments. Our method demonstrates strong generalization results over existing methods on 11 recognition datasets. Here, the CLIP refers to the linear probe CLIP.

Dataset		CLIP	CoOp	CLIP-Adapter	CoCoOp	PLOT	MaPLe	Ours	Δ
Average	Base	69.34	82.69	82.91	80.47	81.3	82.28	84.71	+2.4
	Novel	74.22	63.22	63.98	71.69	72.2	75.14	76.90	+1.8
	HM	71.70	71.66	72.23	75.83	76.48	78.55	80.64	+2.1
ImageNet	Base	72.43	76.47	76.88	75.98	75.33	76.66	77.70	+1.1
	Novel	68.14	67.88	68.1	70.43	70.48	70.54	70.65	+0.1
	HM	70.22	71.92	72.23	73.10	72.83	73.47	74.07	+0.6
Caltech101	Base	96.84	98.00	98.1	97.96	97.86	97.74	98.40	+0.7
	Novel	94.00	89.81	90.00	93.81	93.99	94.36	94.07	-0.3
	HM	95.40	93.73	93.89	95.84	95.92	96.02	96.2	+0.2
OxfordPets	Base	91.17	93.67	93.88	95.20	95.7	95.43	95.67	+0.2
	Novel	97.26	95.29	95.55	97.69	98.1	97.76	97.63	-0.1
	HM	94.12	94.47	94.74	96.43	96.80	96.58	96.67	+0.1
StanfordCars	Base	63.37	78.12	78.35	70.49	71.5	72.94	78.70	+5.8
	Novel	74.89	60.40	60.55	73.59	73.77	74.00	75.67	+1.6
	HM	68.65	68.13	68.33	72.01	72.62	73.47	77.22	+3.8
Flowers102	Base	72.08	97.60	97.61	94.87	95.1	95.92	98.47	+2.5
	Novel	77.80	59.67	59.98	71.75	72.2	72.46	77.00	+4.4
	HM	74.83	74.06	74.32	81.71	82.10	82.56	86.43	+4.0
Food101	Base	90.10	88.33	88.55	90.70	90.98	90.71	91.00	+0.3
	Novel	91.22	82.26	82.35	91.29	91.54	92.05	91.80	-0.9
	HM	90.66	85.19	85.36	90.99	91.28	91.38	91.41	+0.1
FGVCAircraft	Base	27.19	40.44	40.66	33.41	35.6	37.44	43.27	+5.8
	Novel	36.29	22.30	23.1	23.71	28.5	35.61	37.77	+2.0
	HM	31.09	28.75	29.46	27.74	31.66	36.50	40.34	+3.7
SUN397	Base	69.36	80.60	80.85	79.74	79.96	80.82	82.77	+2.0
	Novel	75.35	65.89	65.91	76.86	77.33	78.70	79.07	+0.3
	HM	72.23	72.51	72.62	78.27	78.64	79.75	80.91	+1.2
DTD	Base	53.24	79.44	80.56	77.01	78.9	80.36	83.87	+3.5
	Novel	59.90	41.18	45.30	56.00	57.9	59.18	63.67	+3.5
	HM	56.37	54.24	58	64.85	66.8	68.16	72.20	+4.0
EuroSAT	Base	56.48	92.19	92.5	87.49	90.2	94.07	94.50	+0.5
	Novel	64.05	54.74	55.65	60.04	63.5	73.23	79.60	+6.4
	HM	60.03	68.69	69.49	71.21	74.54	82.35	86.43	+4
UCF101	Base	70.53	84.69	84.10	82.33	82.56	83.00	87.47	+4.5
	Novel	77.50	56.05	57.35	73.45	75.56	78.66	79.17	+0.5
	HM	73.85	67.46	68.21	77.64	78.92	80.77	83.13	+2.5

1.7 Inference Stage Computational Cost

In Table 4, we show the compute cost analysis of our method and compare it with text embedding learning approaches and hand-craft prompt method CLIP-Adapter. We have added experiments on CLIP-Adapter and PLOT.

Table 4: The compute cost comparison using SUN397 dataset. Training time for all methods is calculated for 10 epochs on a single A6000 GPU.

Method	Params	Params % CLIP	Train time (min)	HM
CoOp	2048	0.002	10.88	71.65
CoCoOp	35360	0.03	39.53	75.83
CLIP-Adapter	0.52M	0.41	8.55	72.23
PLOT	8192	0.008	10.85	76.48
MaPLe	3.55 M	2.85	10.58	79.68
Ours	0.65M	0.52	10.21	80.51

1.8 Base-to-novel experiments for ViT-B/16 backbones

Our method demonstrates significant improvements on all 11 datasets. Overall, our method provides the best-averaged results of 84.71%, 76.90%, and 80.64% on the base classes, novel classes, and harmonic mean, respectively. We have added experiments on CLIP-Adapter and PLOT.