

A APPENDIX

This section contains supplementary material that provides additional details for the main paper and further experimental analysis. The content of this section is as follows:

- Additional Experimental Details
- Additional Experimental Analysis
- Additional Ablation Study

A.1 ADDITIONAL EXPERIMENTAL DETAILS

Dataset Details. In Tab. 6, we list the details of the datasets and the hand-crafted prompt we used in the experiments. The prompts are from the Radford et al. (2021) and we have not adopted more prompt templates to generate the optical text representations. In this work, we only focus on the effect of fully fine-tuned CLIP and the text representations would be automatically learned during the training.

Training Details. We maintain the temperature of the softmax function consistent with the pre-trained model, using $\tau = 0.01$, except for when \mathcal{L}_{VLD} is adjusted to 0.1. All images are randomly resized and cropped to 224×224 , only random resize and random crop data augments are applied. The optical hyper-parameter λ is set to 0.7, η is set to 0.1, and α is set to 0.5 for all experiments. We use the AdamW optimizer with the cosine learning rate strategy and the learning rate is set to $5e-6$ and trained for 20 epochs. The batch size is set to 32 for most datasets, with specific batch sizes of 16 for EuroSAT and 64 for ImageNet. For each result of CLIP-CITE, we report the average result with three random seeds.

Table 6: Detailed statistics of the datasets.

Dataset	Classes	Train	Val	Test	Hand-crafted Prompt
Caltech101	100	4,128	1,649	2,465	a photo of a [CLS].
OxfordPets	37	2,944	736	3,669	a photo of a [CLS], a type of pet.
StanfordCars	196	6,509	1,635	8,041	a photo of a [CLS].
Flowers102	102	4,093	1,633	2,463	a photo of a [CLS], a type of flower.
Food101	101	50,500	20,200	30,300	a photo of [CLS], a type of food.
FGVCAircraft	100	3,334	3,333	3,333	a photo of a [CLS], a type of aircraft.
SUN397	397	15,880	3,970	19,850	a photo of a [CLS].
DTD	47	2,820	1,128	1,692	[CLS] texture.
EuroSAT	10	13,500	5,400	8,100	a centered satellite photo of [CLS].
UCF101	101	7,639	1,898	3,783	a photo of a person doing [CLS].
ImageNet	1,000	1.28M	N/A	50,000	a photo of a [CLS]
ImageNetV2	1,000	N/A	N/A	10,000	a photo of a [CLS]
ImageNet-Sketch	1,000	N/A	N/A	50,889	a photo of a [CLS]
ImageNet-A	200	N/A	N/A	7,500	a photo of a [CLS]
ImageNet-R	200	N/A	N/A	30,000	a photo of a [CLS]

A.2 ADDITIONAL EXPERIMENTAL ANALYSIS

Overfitting Analysis. We demonstrate the training process of **FT-Probe** and our **CLIP-CITE** illustrated in Fig. 1 on EuroSAT dataset. The results of loss and accuracy of the training dataset are shown in Fig. 8. We observe that, for the FT-Probe model, there is a decline in the training loss, accompanied by a continual increase in accuracy on the training set. However, the final accuracy on the test set is only 60.86%, which suggests the occurrence of overfitting. In contrast, in the case of our CLIP-CITE model, there is also a reduction in the loss function and a consistent rise in training set accuracy, culminating in a test set accuracy of 95.61%. This indicates that our approach does not exhibit overfitting, demonstrating effectiveness. Moreover, it highlights that overcoming overfitting is a crucial issue when fully fine-tuning models.

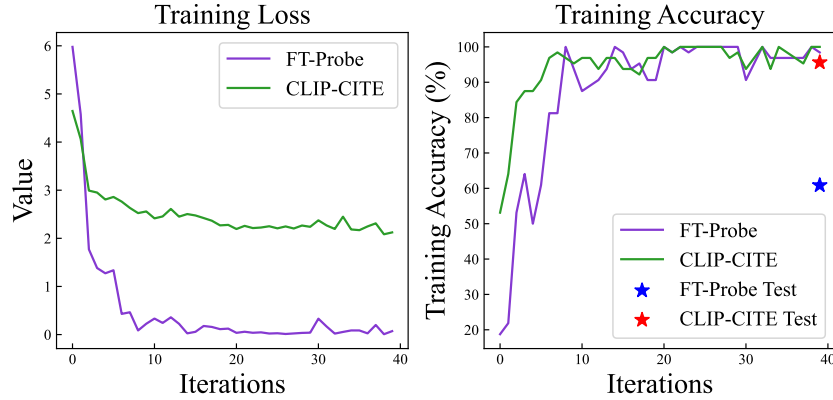


Figure 8: Training loss and accuracy of FT-Probe and CLIP-CITE on EuroSAT dataset.

A.3 ADDITIONAL ABLATION STUDY

Prompt Tuning with Proposed Loss. To evaluate the effectiveness of full-fine-tuning, we also explore the prompt tuning methods with our proposed loss. The results, detailed in Tab. 7, indicate that prompt tuning methods experience a modest improvement with the implementation of our proposed loss functions *i.e.* \mathcal{L}_{SCL} and \mathcal{L}_{VLD} . Notably, our CLIP-CITE still maintains a performance edge. Besides, with the simple fine-tuning (FT-Probe), the tuned model seems to be overfitting, as shown in Fig. 1. Therefore, we propose that both full fine-tuning and well-designed loss functions are crucial in adapting VLMs to the downstream few-shot tasks.

Table 7: Ablation results (%) of our CLIP-CITE and prompt tuning, and fine-tuning methods with various training objectives on the Base-to-New of the ImageNet dataset.

Method	\mathcal{L}_{SCL}	\mathcal{L}_{VLD}	B	N	HM
CLIP			72.43	68.14	70.22
FLYP †			76.21	68.13	71.94
CoOp			76.47	67.88	71.92
CoOp	✓		76.51	67.93	71.97
CoOp	✓	✓	78.23	70.89	72.11
MaPLe			76.66	70.54	73.47
MaPLe	✓		76.70	70.67	73.56
MaPLe	✓	✓	76.71	70.89	73.69
CLIP-CITE	✓	✓	78.44	71.07	74.58

The Effect of the Hyper-Parameter λ and η . In Fig. 9, we ablate the different values on λ and η in Eq. (6). From the results, we observe that the performances in terms of HM are better when applying the \mathcal{L}_{SCL} , *e.g.*, λ is greater than 0. It indicates that supervised vision-language alignment is necessary when fine-tuning. Besides, the vision-language similarity distillation can regularize the model well when η is less than 0.1. In the experiments, the optimal λ and η are set to 0.7 and 0.1, respectively.

Results of Few-Shot Image Recognition. Fig. 10 presents the average results of four competitors and our CLIP-CITE on the 11 datasets under 1, 2, 4, 8, and 16 shots. From the results, we observe that our CLIP-CITE performs very competitively, especially under 1, 2, and 4 shots. When compared with the second-best competitor MaPLe [Khattak et al. \(2023a\)](#) on the average results, our CLIP-CITE demonstrates performance improvements by 3.42%, 3.00%, 2.48%, 1.73%, and 1.52% in scenarios with 1, 2, 4, 8, and 16 shots, respectively. These gains underscore CLIP-CITE’s effectiveness in generalizing to downstream tasks when provided with limited labeled examples. More comparisons of each dataset are provided in the supplementary materials.

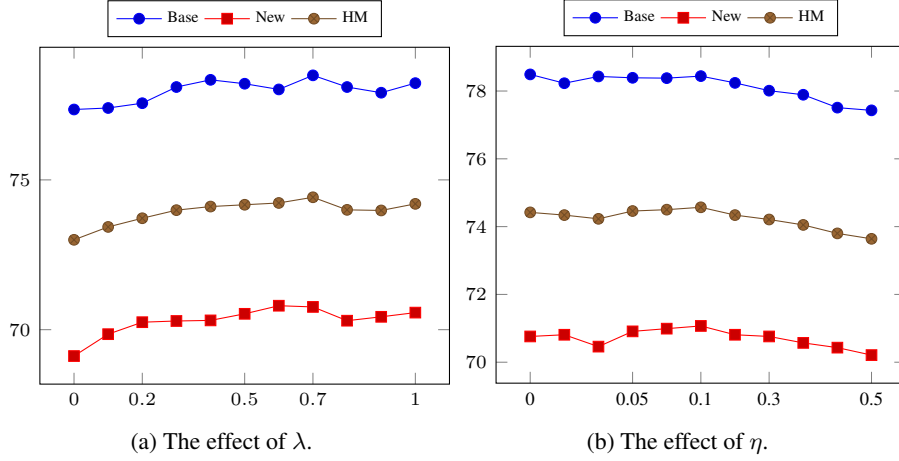


Figure 9: The impacts of the hyper-parameter λ and η on the base-to-new generalization performances. We report the Base (%), New (%), and HM (%) accuracy on the ImageNet dataset.

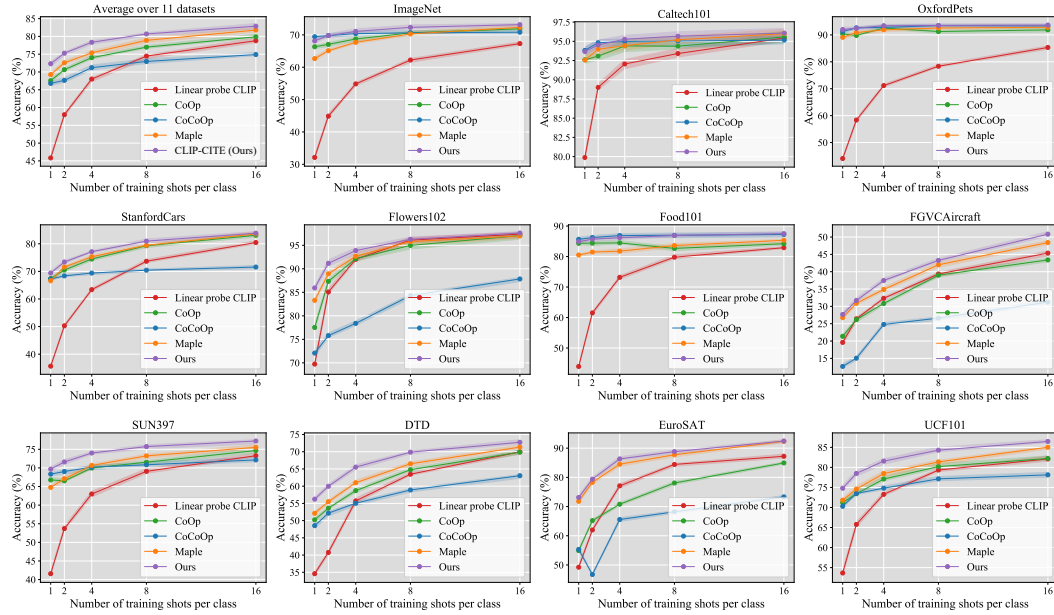


Figure 10: Comparison results of few-shot learning benchmark on the 11 datasets. All of the methods are trained on the ViT-B/16 backbone and implemented with the same experimental settings.

The Effect of Weights Ensemble Ratio α . Tab. 9 shows the results of different datasets with the different ensemble ratios α . Without weights ensemble, CLIP-CITE achieves 85.79%, 73.52%, and 79.19% in Base, New, and HM accuracy, respectively. With the fine-tuning weights ensemble, the performance increases from 71.70% to 78.90% in HM accuracy when α is 0.1. When α increases, the Base accuracy increases, and the New accuracy fluctuates slightly. The optimal value α appears to be 0.5. This indicates that our fine-tuning process maintains a subtle change of model parameters, facilitating smooth compatibility with the zero-shot pre-trained CLIP model and resulting in an overall enhancement of effectiveness.

More Experimental Results of Cross-Domain Generalization Setting. Tab. 8 and Tab. 3 shows the experimental results of Cross-Domain setting. From the results of Tab. 8, all methods trained on the ImageNet can consistently obtain the generalization performance on the other 10 datasets. From the results of Tab. 3, the prompt tuning methods trained on other datasets are difficult to transfer to ImageNet and impact the overall generalization, while our fine-tuning methods can maintain or even

Methods	CLIP		CoOp		CoCoOp		MaPLe		CLIP-CITE	
	Base	New	Base	New	Base	New	Base	New	Base	New
Caltech101	96.84	94.00	94.15	93.92	96.58	95.16	96.30	94.98	96.71	93.82
OxfordPets	91.17	97.26	90.34	97.69	90.8	97.97	90.57	97.73	89.56	96.78
StanfordCars	63.37	74.89	61.99	73.37	63.62	74.48	62.44	73.98	60.74	72.41
Flowers102	72.08	77.80	66.86	75.23	72.30	77.64	73.25	76.86	71.48	76.9
Food101	90.10	91.22	88.62	90.68	89.39	91.0	89.29	90.86	88.47	90.53
FGVCAircraft	27.19	36.29	21.21	26.36	27.65	32.37	28.69	31.21	26.33	34.33
SUN397	69.36	75.35	68.36	72.78	72.08	75.96	71.46	76.1	71.78	76.16
DTD	53.24	59.90	49.00	51.73	55.32	57.01	51.04	54.51	50.39	57.53
EuroSAT	56.48	64.05	50.2	69.22	52.1	68.84	47.52	59.83	49.50	65.51
UCF101	70.53	77.50	68.89	71.88	70.89	75.77	69.22	74.97	70.99	76.55

Table 8: Cross-Domain evaluation. All the models are trained on the base training set of the ImageNet dataset and evaluated on the 10 datasets .

α ratio		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Average on	Base	69.34	83.53	84.66	84.44	85.09	85.48	85.64	85.69	85.64	85.58	85.79
	New	74.22	74.75	76.07	75.71	75.74	77.08	74.68	75.58	75.62	75.77	73.52
	HM	71.70	78.90	80.13	79.83	80.15	81.06	79.79	80.32	80.32	80.38	79.19
ImageNet	Base	72.43	77.45	77.63	78.20	78.23	78.44	78.44	78.48	78.46	78.49	78.50
	New	68.14	70.35	70.79	70.71	70.65	71.07	70.32	70.59	70.36	70.29	70.23
	HM	70.22	73.73	74.05	74.27	74.25	74.58	74.16	74.33	74.19	74.17	74.14
Caltech101	Base	96.84	97.20	97.65	97.78	98.77	98.82	98.82	98.83	98.83	98.83	98.85
	New	94.00	93.40	94.14	93.44	93.65	94.28	93.53	93.90	94.00	93.47	93.20
	HM	95.40	95.26	95.86	95.56	96.14	96.50	96.10	96.30	96.36	96.08	95.94
OxfordPets	Base	91.17	95.23	95.84	95.66	95.82	96.01	96.18	96.42	96.60	96.93	97.01
	New	97.26	96.12	96.47	96.69	96.71	97.95	96.66	96.72	96.90	97.28	95.23
	HM	94.12	95.67	96.15	96.17	96.27	96.97	96.42	96.57	96.75	97.11	96.11

Table 9: Comparison with the different ensemble ratio α on base-to-new generalization.

enhance the performance of ImageNet. These demonstrate that ImageNet encompasses a broader array of patterns and categories, and both prompt tuning methods and our approach effectively sustain performance across various datasets. When transferring from other datasets to ImageNet, CLIP-CITE can uphold ImageNet’s performance. It shows that our fine-tuning method has better generalization capacity.

Dataset	Method	1-shot	2-shot	4-shot	8-shot	16-shot
Average	Linear probe CLIP	45.83	57.98	68.01	74.47	78.79
	CoOp	67.56	70.65	74.02	76.98	79.89
	CoCoOp	66.79	67.65	71.21	72.96	74.90
	MaPLe	69.27	72.58	75.37	78.89	81.79
	CLIP-CITE	72.69	75.58	77.85	80.62	83.31
ImageNet	Linear probe CLIP	32.13	44.88	54.85	62.23	67.31
	CoOp	66.33	67.07	68.73	70.63	71.87
	CoCoOp	69.43	69.78	70.39	70.63	70.83
	MaPLe	62.67	65.10	67.70	70.30	72.33
	CLIP-CITE	68.20	68.90	70.30	71.20	72.90
Caltech101	Linear probe CLIP	79.88	89.01	92.05	93.41	95.43
	CoOp	92.60	93.07	94.4	94.37	95.57
	CoCoOp	93.83	94.82	94.98	95.04	95.16
	MaPLe	92.57	93.97	94.43	95.2	96.00
	CLIP-CITE	94.16	94.81	95.53	96.39	96.50
OxfordPets	Linear probe CLIP	44.06	58.37	71.17	78.36	85.34
	CoOp	90.37	89.8	92.57	91.27	91.87
	CoCoOp	91.27	92.64	92.81	93.45	93.34
	MaPLe	89.10	90.87	91.9	92.57	92.83
	CLIP-CITE	91.47	93.02	93.54	93.87	94.70
StanfordCars	Linear probe CLIP	35.66	50.28	63.38	73.67	80.44
	CoOp	67.43	70.5	74.47	79.3	83.07
	CoCoOp	67.22	68.37	69.39	70.44	71.57
	MaPLe	66.60	71.60	75.30	79.47	83.57
	CLIP-CITE	70.63	74.22	76.53	79.94	83.70
Food101	Linear probe CLIP	43.96	61.51	73.19	79.79	82.90
	CoOp	84.33	84.40	84.47	82.67	84.20
	CoCoOp	85.65	86.22	86.88	86.97	87.25
	MaPLe	80.50	81.47	81.77	83.60	85.33
	CLIP-CITE	85.16	85.95	86.05	86.68	87.00
Flowers102	Linear probe CLIP	69.74	85.07	92.02	96.10	97.37
	CoOp	77.53	87.33	92.17	94.97	97.07
	CoCoOp	72.08	75.79	78.40	84.30	87.84
	MaPLe	83.30	88.93	92.67	95.80	97.00
	CLIP-CITE	84.25	86.76	92.08	95.86	97.6
FGVCAircraft	Linear probe CLIP	19.61	26.41	32.33	39.35	45.36
	CoOp	21.37	26.20	30.83	39.00	43.40
	CoCoOp	12.68	15.06	24.79	26.61	31.21
	MaPLe	26.73	30.90	34.87	42.00	48.40
	CLIP-CITE	29.34	32.40	36.60	46.00	57.00
SUN397	Linear probe CLIP	41.58	53.70	63.00	69.08	73.28
	CoOp	66.77	66.53	69.97	71.53	74.67
	CoCoOp	68.33	69.03	70.21	70.84	72.15
	MaPLe	64.77	67.10	70.67	73.23	75.53
	CLIP-CITE	69.54	70.99	72.36	74.45	76.30
DTD	Linear probe CLIP	34.59	40.76	55.71	63.46	69.96
	CoOp	50.23	53.60	58.70	64.77	69.87
	CoCoOp	48.54	52.17	55.04	58.89	63.04
	MaPLe	52.13	55.50	61.00	66.50	71.33
	CLIP-CITE	54.20	60.70	64.54	67.67	72.50
EuroSAT	Linear probe CLIP	49.23	61.98	77.09	84.43	87.21
	CoOp	54.93	65.17	70.80	78.07	84.93
	CoCoOp	55.33	46.74	65.56	68.21	73.32
	MaPLe	71.80	78.30	84.50	87.73	92.33
	CLIP-CITE	76.20	85.20	88.77	91.17	92.60
UCF101	Linear probe CLIP	53.66	65.78	73.28	79.34	82.11
	CoOp	71.23	73.43	77.10	80.20	82.23
	CoCoOp	70.30	73.51	74.82	77.14	78.14
	MaPLe	71.83	74.60	78.47	81.37	85.03
	CLIP-CITE	76.40	78.38	80.07	83.56	85.70

Table 10: Per-dataset performance comparison of our method with various methods in the few-shot setting.