Style-Pro: Style-Guided Prompt Learning for Generalizable Vision-Language Models

Niloufar Alipour Talemi, Hossein Kashiani, Fatemeh Afghah

Clemson University

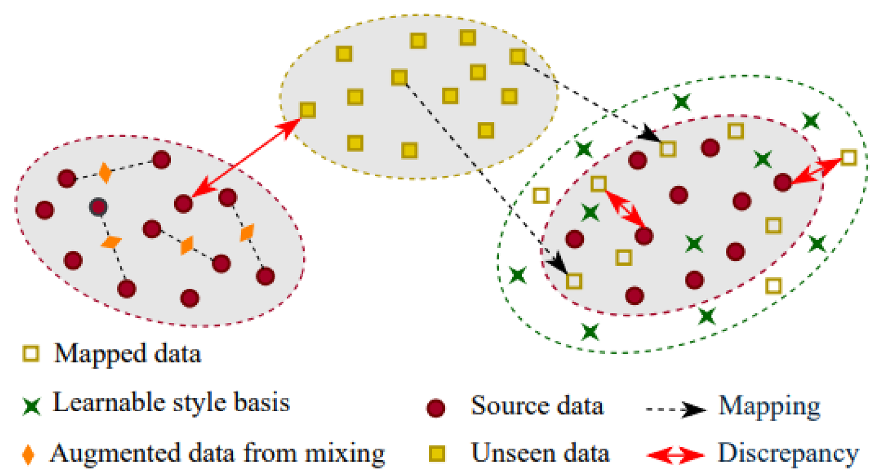
Holcombe Department of Electrical and Computer Engineering, Clemson University

WACV 2025 · TUCSON, ARIZONA • FEB 28 - MAR 4

Challenges

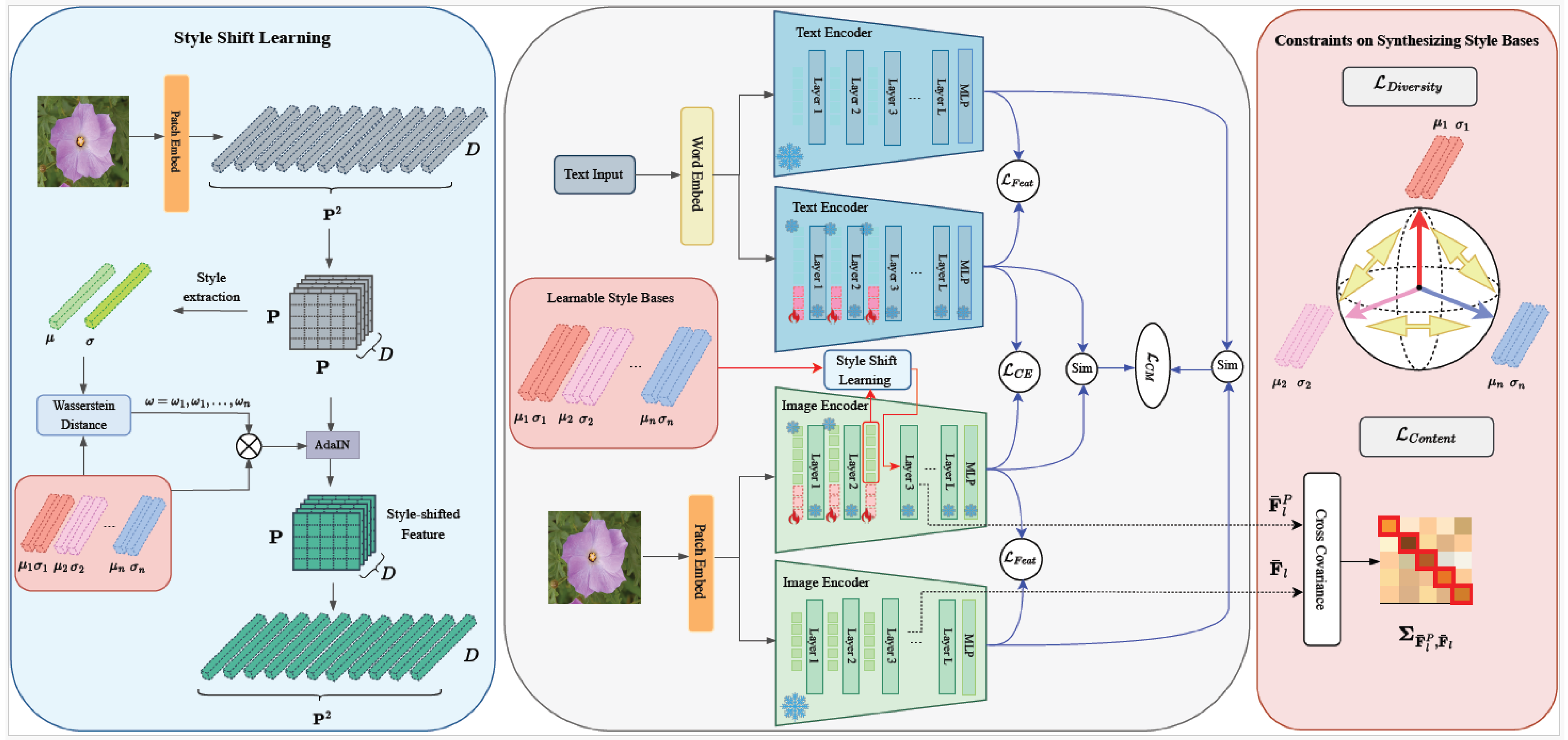
- Few-shot fine-tuning often leads to overfitting when optimizing prompts for task-specific objectives, restricting the model's ability to generalize beyond the training samples.
- This overfitting poses a significant challenge for adapting vision-language models to new domains or unseen classes within the same domain.

Contributions



- We propose Style-Pro, a novel style guided prompt learning framework that mitigates overfitting and preserves the zero-shot generalization capabilities of CLIP.
- Style-Pro employs learnable style bases to synthesize styles beyond the source domain and maps unseen styles as weighted combinations to reduce domain discrepancies.
- Extensive experiments across 11 benchmark datasets demonstrate the effectiveness of Style-Pro, consistently surpassing state-of-the-art methods in various settings, including base-to-new generalization, cross-dataset transfer, and domain generalization.

Style-Pro Framework



Methodology

The proposed Style-Pro framework integrates two complementary strategies: style shift learning and consistency constraints.

- **Style shift learning**: Consider a collection of style base: $B_{sty} = (\mu_b^n, \sigma_b^n)_{n=1}^N$

Compute the Wasserstein distance to determine the discrepancy in style distribution:

$$d_{cur} = ||\mu_{cur} - \mu_b^n||_2^2 + (\sigma_{cur}^2 + \sigma_b^{n\,2} - 2\sigma_{cur}\sigma_b^n)$$

Map the unseen styles into the known style representation space:

$$\mu_{map} = \sum_{n=1}^{N} \omega_n \mu_b^n, \quad \sigma_{map} = \sum_{n=1}^{N} \omega_n \sigma_b^n \qquad \mathbf{F}_l'' = \sigma_{map}\bar{\mathbf{F}}_l' + \mu_{map}$$

- **Self-consistency regularization**: Ensures the prompted model maintains its generalization capability across new classes and diverse domains.

Feature-level Alignment

$$\mathcal{L}_{Feat} = \frac{1}{d}\left(\lambda_f \sum_{i=1}^{d} (\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_{p_i})^2 + \lambda_g \sum_{i=1}^{d} (\tilde{\mathbf{g}}_i - \tilde{\mathbf{g}}_{p_i})^2\right)$$

Cross-modality Alignment

$$\mathcal{L}_{CM} = \mathcal{D}_{\mathrm{KL}}(Pre, Pre_p)$$
$$Pre = \mathrm{sim}(\tilde{f}, \tilde{g}), \; Pre_p = \mathrm{sim}(\tilde{f}_p, \tilde{g}_p)$$

Experimental Results

Base-to-novel generalization evaluation.

Dataset		CLIP [56]	CoOp [87]	CoCoOp [86]	MaPLe [38]	PromptSRC [39]	CoPrompt [62]	MMA [79]	Style-Pro (Proposed)
Average on 11 datasets	B	69.34	82.69	80.47	82.28	84.26	84.00	83.20	**84.48**
	N	74.22	63.22	71.69	75.14	76.10	77.23	76.80	**78.06**
	H	71.70	71.66	75.83	78.55	79.97	80.48	79.87	**80.98**
ImageNet	B	72.43	76.47	75.98	76.66	77.60	**77.67**	77.31	77.58
	N	68.14	67.88	70.43	70.54	70.73	71.27	71.00	**71.68**
	H	70.22	71.92	73.10	73.47	74.01	74.33	74.02	**74.51**
Caltech101	B	96.84	98.00	97.96	97.74	98.10	98.27	**98.40**	98.38
	N	94.00	89.81	93.81	94.36	94.03	94.90	94.00	**95.44**
	H	95.40	93.73	95.84	96.02	96.02	96.55	96.15	**96.89**
OxfordPets	B	91.17	93.67	95.20	95.43	95.33	**95.67**	95.40	95.64
	N	97.26	95.29	97.69	97.76	97.30	98.10	98.07	**98.63**
	H	94.12	94.47	96.43	96.58	96.30	96.87	96.72	**97.11**
Stanford Cars	B	63.37	78.12	70.49	72.94	78.27	76.97	78.50	**78.53**
	N	74.89	60.40	73.59	74.00	74.97	74.40	73.10	**75.12**
	H	68.65	68.13	72.01	73.47	76.58	75.66	75.70	**76.79**
Flowers 102	B	72.08	97.60	94.87	95.92	**98.07**	97.27	97.77	98.04
	N	77.80	59.67	71.75	72.46	76.50	76.60	75.93	**76.86**
	H	74.83	74.06	81.71	82.56	85.95	85.71	85.48	**86.17**
Food101	B	92.43	88.33	90.70	90.71	90.67	90.73	90.13	**90.93**
	N	91.22	82.26	91.29	92.05	91.53	92.07	91.30	**92.29**
	H	90.66	85.19	90.99	91.38	91.10	91.40	90.71	**91.60**
FGVC Aircraft	B	27.19	40.44	33.41	37.44	42.73	40.20	40.57	**42.79**
	N	36.29	22.30	23.71	35.61	37.87	**39.33**	36.33	39.28
	H	31.09	28.75	27.74	36.50	40.15	39.76	38.33	**40.96**
SUN397	B	69.36	80.60	79.74	80.82	**82.67**	82.63	82.27	82.66
	N	75.35	65.89	76.86	78.70	78.47	80.03	78.57	**80.61**
	H	72.23	72.51	78.27	79.75	80.52	81.31	80.38	**81.62**
DTD	B	53.24	79.44	77.01	80.36	83.37	83.13	83.20	**83.41**
	N	59.90	41.18	56.00	59.18	62.97	64.73	**65.63**	65.58
	H	56.37	54.24	64.85	68.16	71.75	72.79	73.38	**73.43**
EuroSAT	B	56.48	92.19	87.49	94.07	92.90	**94.60**	85.46	94.52
	N	64.05	54.74	60.04	73.23	73.90	78.57	82.34	**82.74**
	H	60.03	68.69	71.21	82.35	82.32	85.84	83.87	**88.24**
UCF101	B	70.53	84.69	82.33	83.00	87.10	**86.90**	86.23	86.83
	N	77.50	56.05	73.45	78.66	78.80	79.57	80.03	**80.40**
	H	73.85	67.46	77.64	80.77	82.74	83.07	82.20	**83.49**

Cross-dataset evaluation.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoOp [87]	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp [86]	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe [38]	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromtSCR [39]	**71.27**	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
CoPrompt [62]	70.80	94.50	90.73	65.67	72.30	86.43	24.00	67.57	47.07	**51.90**	69.73	67.00
MMA [79]	71.00	93.80	90.30	**66.13**	72.07	86.12	**25.33**	68.17	46.57	49.24	68.32	66.61
Style-Pro	71.23	**94.66**	**90.91**	66.03	**72.54**	**86.61**	25.14	**68.38**	**47.29**	50.85	**69.96**	**67.24**

Domain generalization evaluation.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP [56]	66.73	60.83	46.15	47.77	73.96	57.18
CoOp [87]	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp [86]	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe [38]	70.72	64.07	49.15	50.90	76.98	60.27
PromptSRC [39]	**71.27**	64.35	49.55	50.90	77.80	60.65
CoPrompt [62]	70.80	64.81	49.54	51.51	77.34	60.80
MMA [79]	71.00	64.33	49.13	51.12	77.32	60.77
Style-Pro	71.23	**65.66**	**50.38**	**51.93**	**77.98**	**61.49**

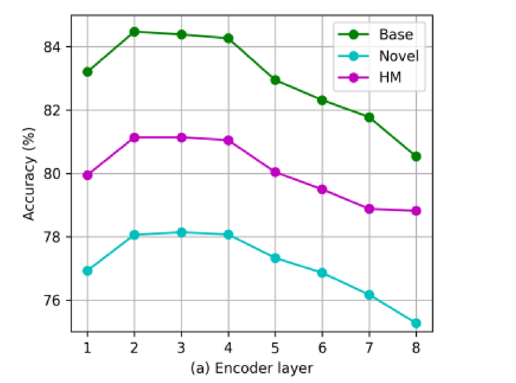
Experiments

- We validate our method across three different settings: generalization from base-to-novel, classes, cross-dataset evaluation, and domain generalization.
- For base-to-novel and cross-dataset experiments: Generic-object datasets (ImageNet and Caltech101), Fine-grained datasets (Oxford Pets, Stanford Cars, Flowers102, Food101, and FGVC Aircraft), remote sensing classification dataset (EuroSAT), scene recognition dataset (SUN397), Action recognition dataset (UCF101), Texture dataset (DTD). For domain generalization experiments: ImageNetV2, ImageNet Sketch, ImageNet-A, ImageNet-R.
- Ablations studies proves that components complement each other to mitigate overfitting in vision-language model adaptation, leading to improved generalization performance.

Analysis of different constraints of Style-Pro framework.

Approach				Accuracy		
Consistency		Style Shift		Base	Novel	HM
Feat	CM	Content	Diversity			
				82.51	73.36	77.66
✓				82.77	74.28	78.30
✓	✓			82.97	75.64	79.14
✓	✓	✓		83.11	76.09	79.45
✓	✓	✓	✓	**84.48**	**78.06**	**80.98**

Ablation study on style shift learning at different layers of the vision encoder.



Ablation study on the impact of the number of learnable style bases.