PromptStyler: Prompt-driven Style Generation for Source-free Domain Generalization

Junhyeong Cho[1] Gilhyun Nam[1] Sungyeon Kim[2] Hunmin Yang[1,3] Suha Kwak[2]
[1]ADD [2]POSTECH [3]KAIST

ICCV 2023

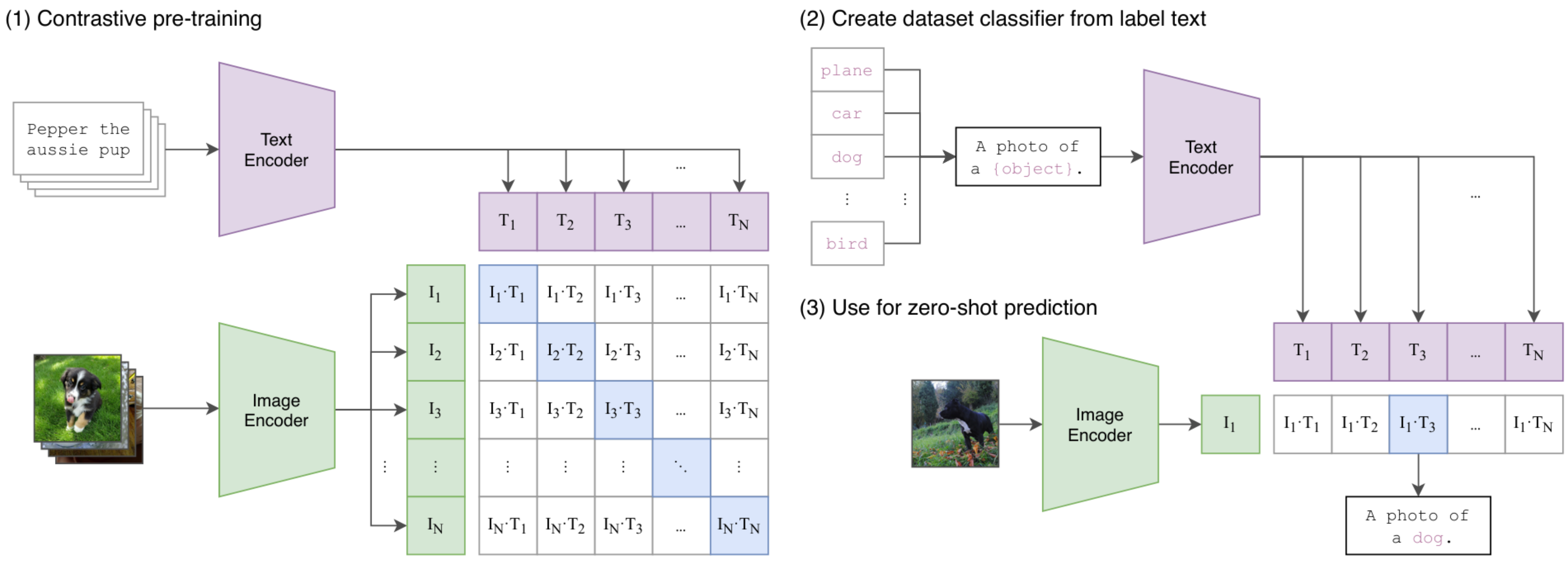
Background

Compararions between different setup

Setup	Source	Target	Task Definition
DA	✓	✓	✓
DG	✓	–	✓
Source-free DA	–	✓	✓
Source-free DG	–	–	✓

Table 1: Different requirements in each setup. Source-free DG only assumes the task definition (*i.e.*, what should be predicted) without requiring source and target domain data.

Background



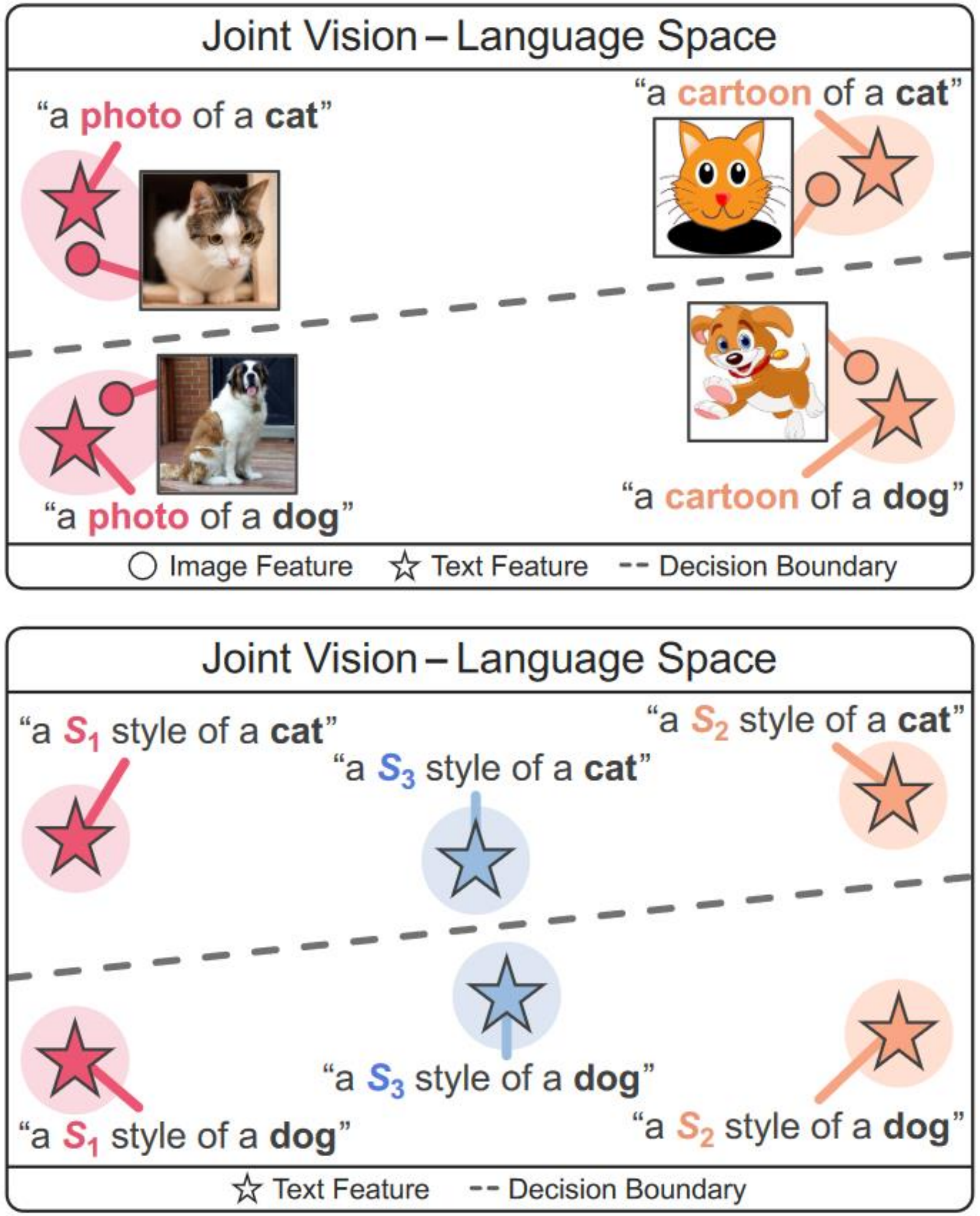
(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Motivation

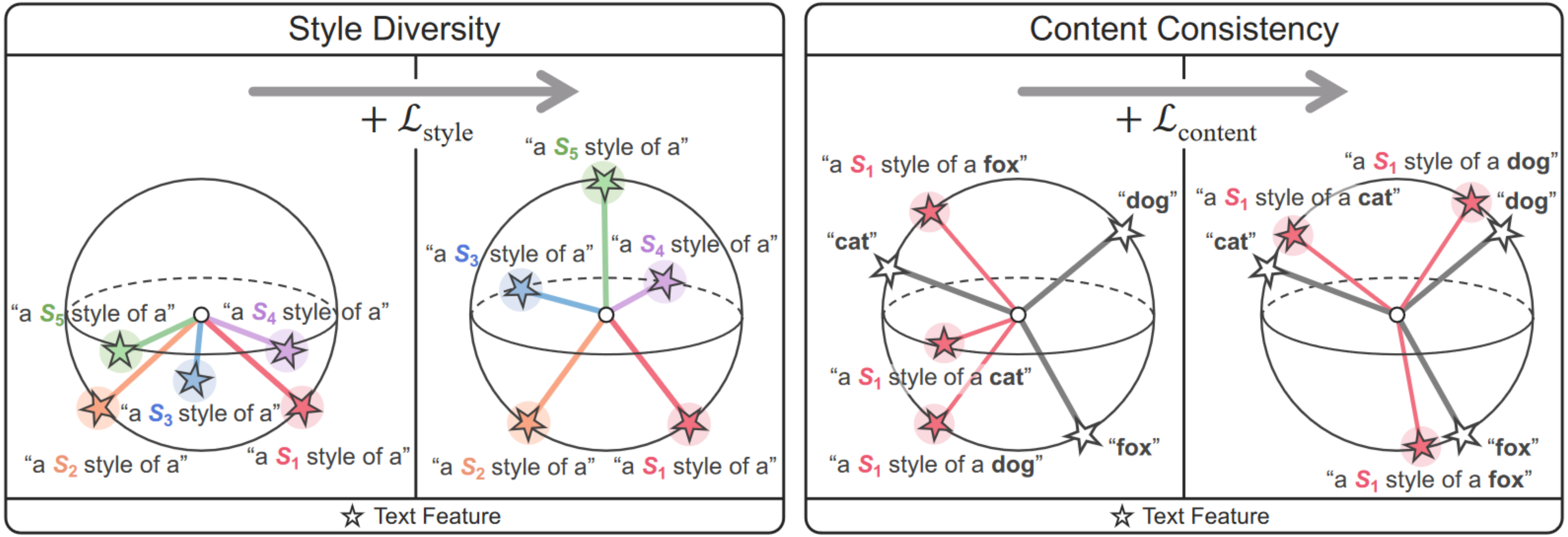
- Text features could effectively represent various image styles in a joint vision-language space.

- PromptStyler synthesizes diverse styles in a joint vision-language space via learnable style word vectors for pseudo-words S* without using any images.

(a)
(b)

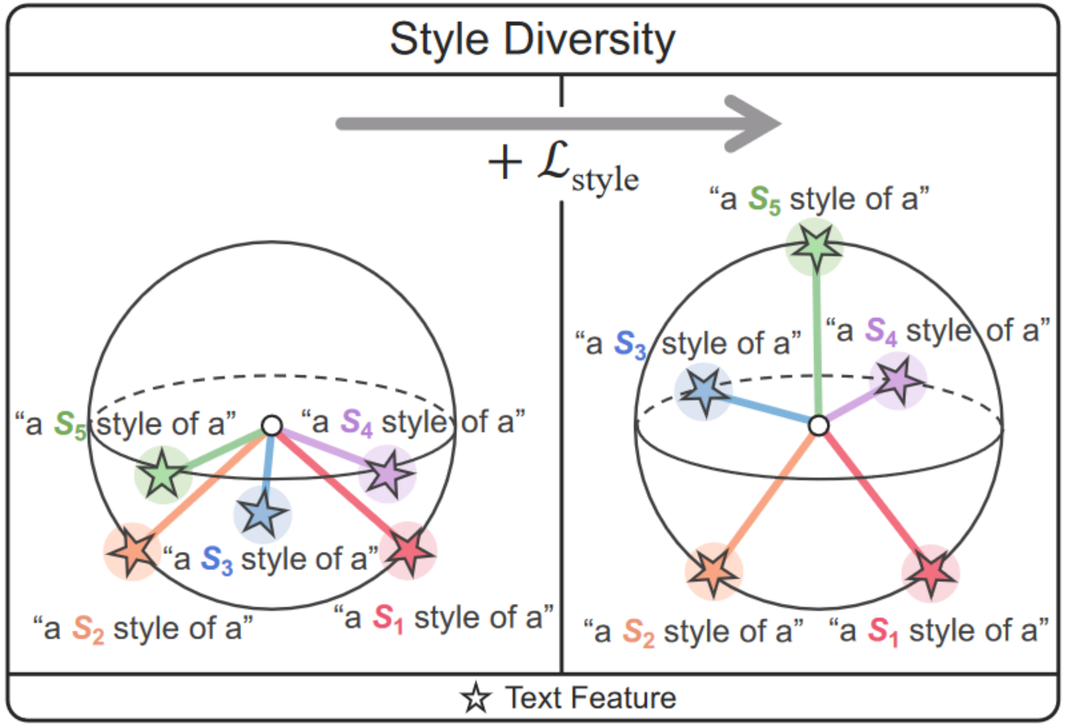


Methods



PromptStyler: A prompt-driven style generation method
To simulate distribution shifts

Methods



Prompt-driven style generation

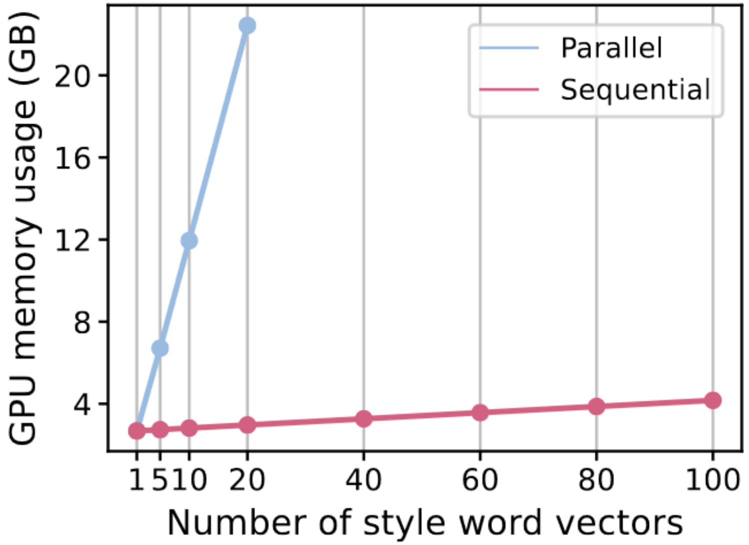
P_i^{style} a S_i style of a

$P_m^{content}$ $[class]_m$

$P_i^{style} \circ P_m^{content}$ a S_i style of a$[class]_m$

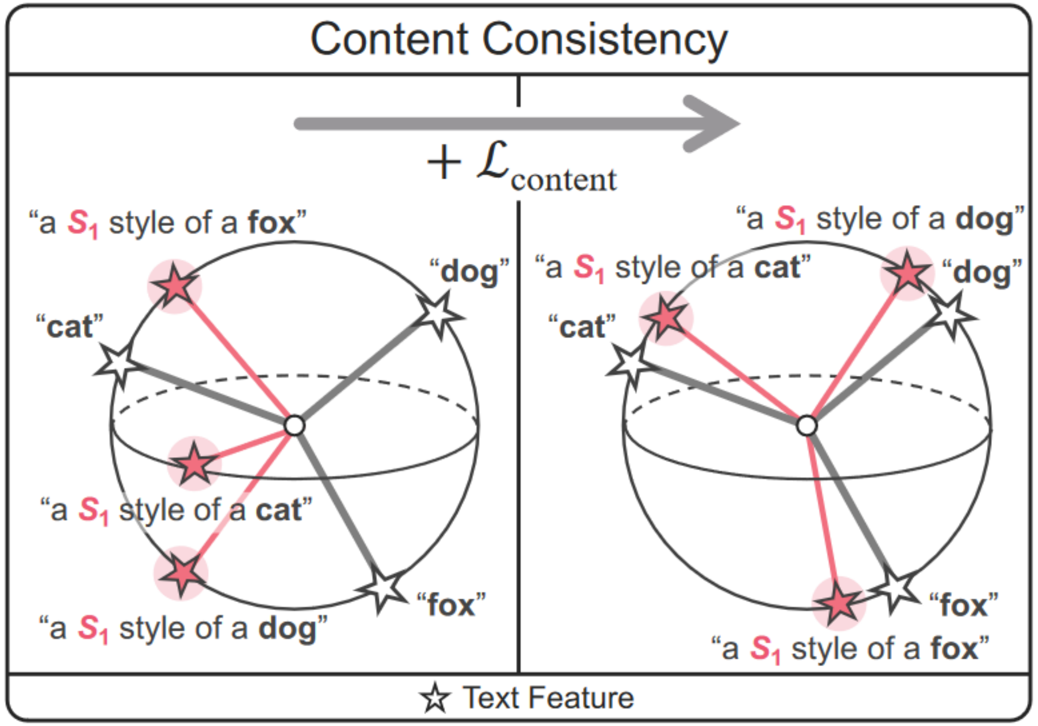
To learn K style word vectors $\{s_i\}_{i=1}^{K}$

- Sequentially learn style word vectors
- Feature of s_i orthogonal to all previous(1,2,…,i-1)



Style diversity loss

$$\mathcal{L}_{\text{style}} = \frac{1}{i-1} \sum_{j=1}^{i-1} \left| \frac{T(\mathcal{P}_i^{\text{style}})}{\|T(\mathcal{P}_i^{\text{style}})\|_2} \cdot \frac{T(\mathcal{P}_j^{\text{style}})}{\|T(\mathcal{P}_j^{\text{style}})\|_2} \right| .$$



Content consistency loss

Define a cosine similarity score z_{imn} as

$$z_{imn} = \frac{T(\mathcal{P}_i^{\text{style}} \circ \mathcal{P}_m^{\text{content}})}{\|T(\mathcal{P}_i^{\text{style}} \circ \mathcal{P}_m^{\text{content}})\|_2} \cdot \frac{T(\mathcal{P}_n^{\text{content}})}{\|T(\mathcal{P}_n^{\text{content}})\|_2} .$$

Calculate the content consistency loss as

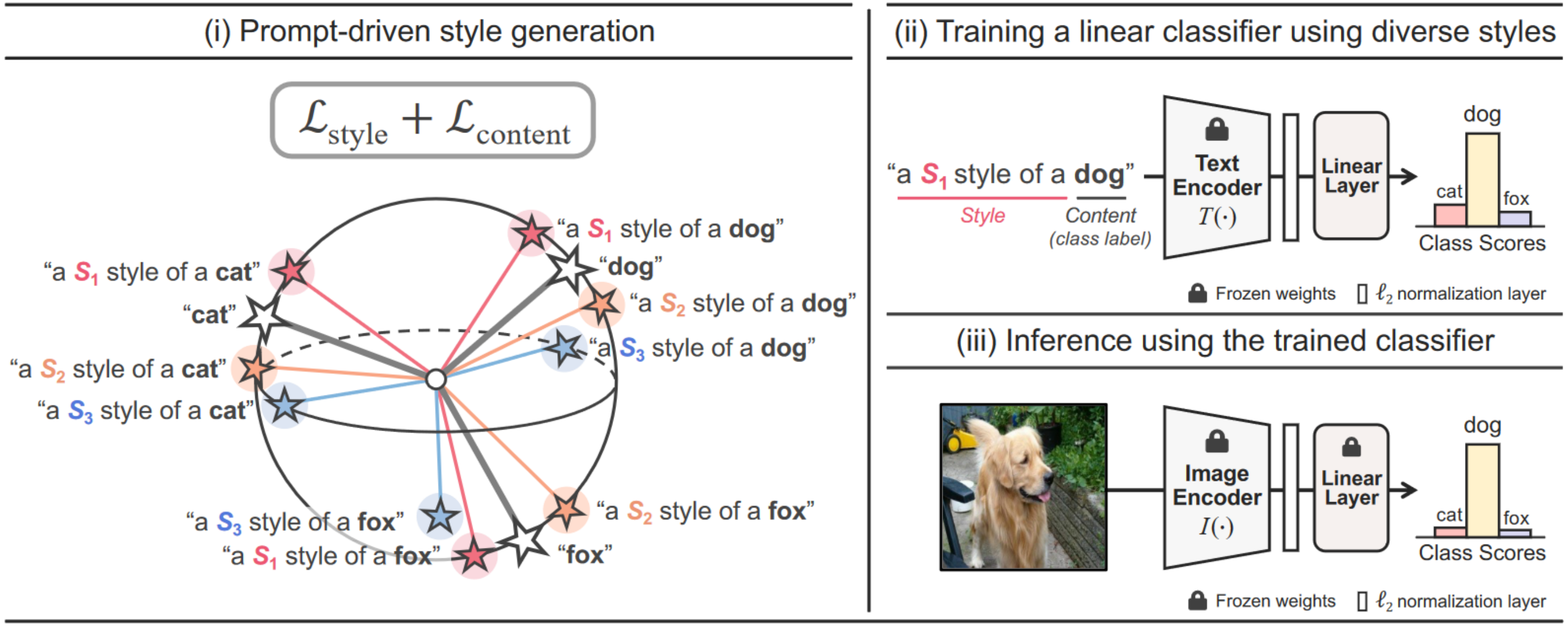
$$\mathcal{L}_{\text{content}} = -\frac{1}{N}\sum_{m=1}^{N} \log\left(\frac{\exp(z_{imm})}{\sum_{n=1}^{N} \exp(z_{imn})}\right),$$

Total prompt loss

$$\mathcal{L}_{\text{prompt}} = \mathcal{L}_{\text{style}} + \mathcal{L}_{\text{content}} .$$

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



Method	Inference Module: Image Encoder	Inference Module: Text Encoder	# Params	FPS
OfficeHome (65 classes)				
ZS-CLIP [50]	✓	✓	102.0M	1.6
PromptStyler	✓	–	**38.4M**	**72.9**
DomainNet (345 classes)				
ZS-CLIP [50]	✓	✓	102.0M	0.3
PromptStyler	✓	–	**38.7M**	**72.9**

Experiments

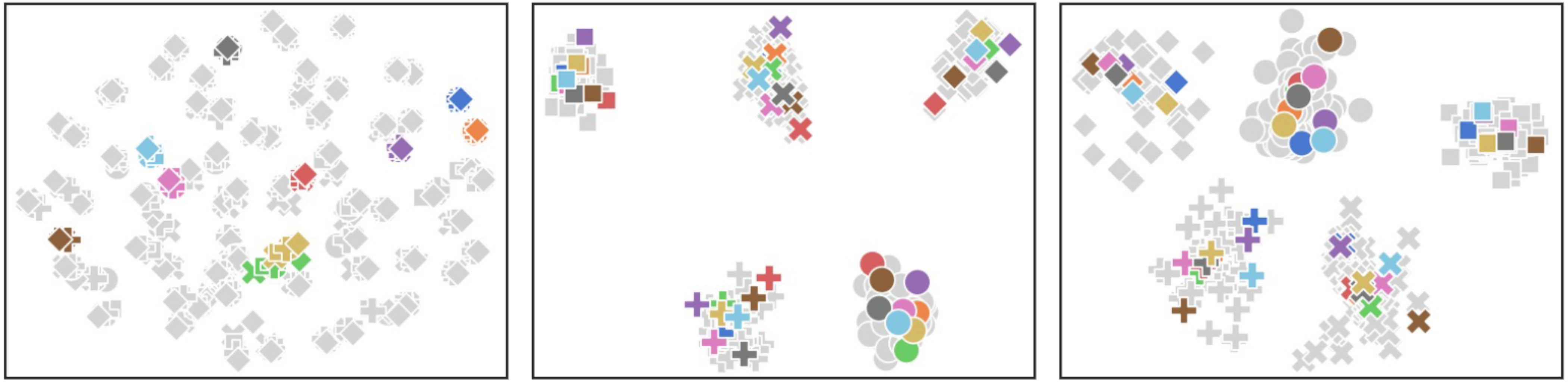
Method	Configuration: Source Domain	Configuration: Domain Description	Accuracy (%): PACS	VLCS	OfficeHome	DomainNet	Avg.
ResNet-50 [22] with pre-trained weights on ImageNet [6]							
DANN [19]	✓	–	83.6±0.4	78.6±0.4	65.9±0.6	38.3±0.1	66.6
RSC [25]	✓	–	85.2±0.9	77.1±0.5	65.5±0.9	38.9±0.5	66.7
MLDG [35]	✓	–	84.9±1.0	77.2±0.4	66.8±0.6	41.2±0.1	67.5
SagNet [46]	✓	–	**86.3**±0.2	77.8±0.5	68.1±0.1	40.3±0.1	68.1
SelfReg [28]	✓	–	85.6±0.4	77.8±0.9	67.9±0.7	42.8±0.0	68.5
GVRT [44]	✓	–	85.1±0.3	**79.0**±0.2	70.1±0.1	44.1±0.1	69.6
MIRO [5]	✓	–	85.4±0.4	**79.0**±0.0	**70.5**±0.4	**44.3**±0.2	**69.8**
ResNet-50 [22] with pre-trained weights from CLIP [50]							
ZS-CLIP (C) [50]	–	–	90.6±0.0	76.0±0.0	68.6±0.0	45.6±0.0	70.2
CAD [53]	✓	–	90.0±0.6	81.2±0.6	70.5±0.3	45.5±2.1	71.8
ZS-CLIP (PC) [50]	–	✓	90.7±0.0	80.1±0.0	72.0±0.0	46.2±0.0	72.3
PromptStyler	–	–	**93.2**±0.0	**82.3**±0.1	**73.6**±0.1	**49.5**±0.0	**74.7**
ViT-B / 16 [11] with pre-trained weights from CLIP [50]							
ZS-CLIP (C) [50]	–	–	95.7±0.0	76.4±0.0	79.9±0.0	57.8±0.0	77.5
MIRO [5]	✓	–	95.6	82.2	82.5	54.0	78.6
ZS-CLIP (PC) [50]	–	✓	96.1±0.0	82.4±0.0	82.3±0.0	57.7±0.0	79.6
PromptStyler	–	–	**97.2**±0.1	**82.9**±0.0	**83.6**±0.0	**59.4**±0.0	**80.8**
ViT-L / 14 [11] with pre-trained weights from CLIP [50]							
ZS-CLIP (C) [50]	–	–	97.6±0.0	77.5±0.0	85.9±0.0	63.3±0.0	81.1
ZS-CLIP (PC) [50]	–	✓	98.5±0.0	**82.4**±0.0	86.9±0.0	64.0±0.0	83.0
PromptStyler	–	–	**98.6**±0.0	**82.4**±0.2	**89.1**±0.0	**65.5**±0.0	**83.9**

Experiments

$\mathcal{L}_{\text{style}}$	$\mathcal{L}_{\text{content}}$	Accuracy (%)				
		PACS	VLCS	OfficeHome	DomainNet	Avg.
–	–	92.6	78.3	72.2	48.0	72.8
✓	–	92.3	80.9	71.5	48.2	73.2
–	✓	92.8	80.5	72.4	48.6	73.6
✓	✓	**93.2**	**82.3**	**73.6**	**49.5**	**74.7**

Table 4: Ablation study on the style diversity loss $\mathcal{L}_{\text{style}}$ and content consistency loss $\mathcal{L}_{\text{content}}$ used in the prompt loss.

Experiments



(a) $\mathcal{L}_{\text{style}}$ (b) $\mathcal{L}_{\text{content}}$ (c) $\mathcal{L}_{\text{style}} + \mathcal{L}_{\text{content}}$

Figure 5: Text-to-Image synthesis results using style-content features (from "a S_* style of a **cat**") with 6 different style word vectors. By leveraging the proposed method, we could learn a variety of styles while not distorting content information.

Experiments

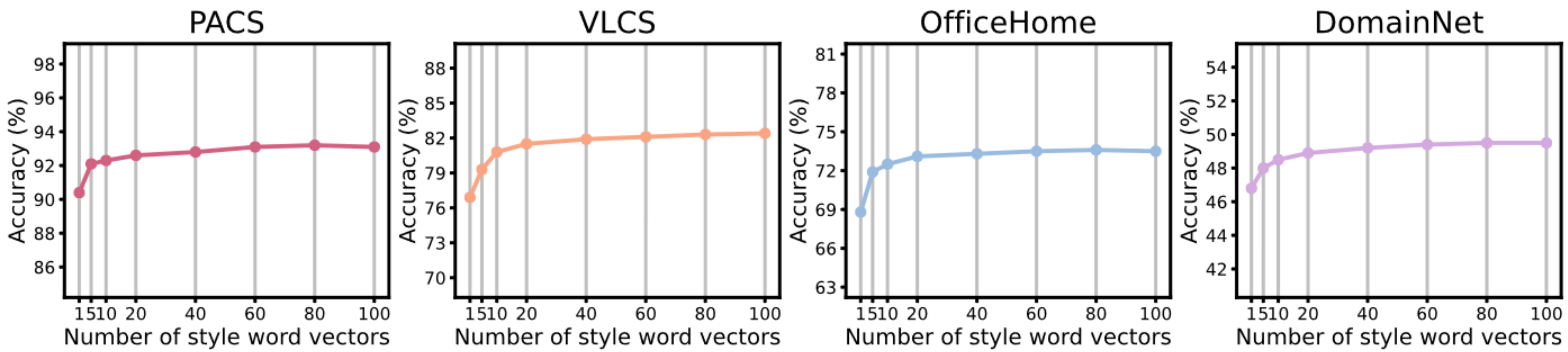


Figure 6: Top-1 classification accuracy on the PACS [34], VLCS [15], OfficeHome [60] and DomainNet [48] datasets with regard to the number of learnable style word vectors K.

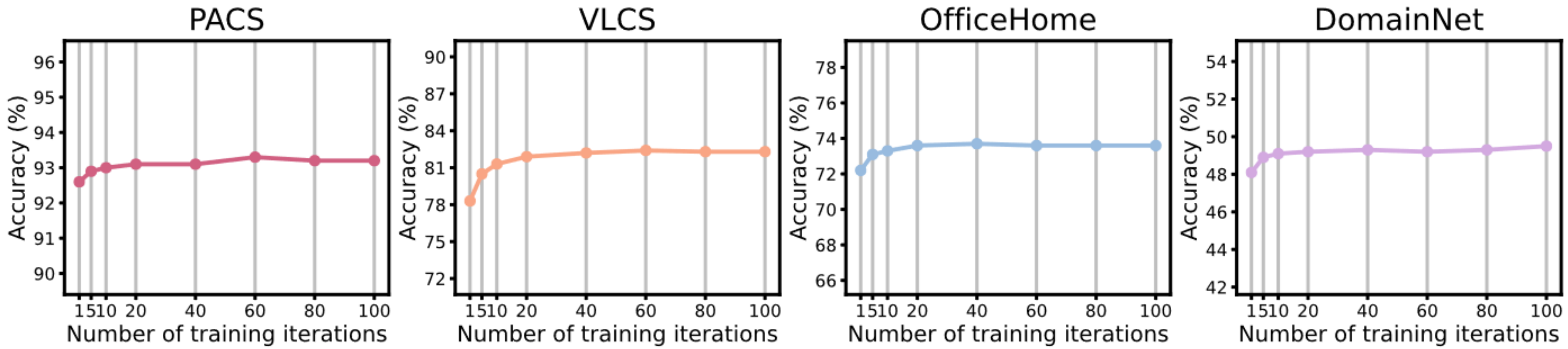


Figure 7: Top-1 classification accuracy on the PACS [34], VLCS [15], OfficeHome [60] and DomainNet [48] datasets with regard to the number of training iterations L for learning each style word vector s_i.

Experiments

(a) Dog (b) Cat (c) Squirrel

Figure B1: Several examples from the Terra Incognita [1] dataset. We visualize class entities using red bounding boxes, since they are not easily recognizable due to their small sizes and complex background scenes.

Method	Configuration: Source Domain	Configuration: Domain Description	Accuracy (%): Location100	Location38	Location43	Location46	Avg.
ResNet-50 [22] with pre-trained weights on ImageNet [6]							
SelfReg [28]	✓	–	48.8±0.9	41.3±1.8	57.3±0.7	**40.6**±0.9	47.0
GVRT [44]	✓	–	**53.9**±1.3	**41.8**±1.2	**58.2**±0.9	38.0±0.6	**48.0**
ResNet-50 [22] with pre-trained weights from CLIP [50]							
ZS-CLIP (C) [50]	–	–	8.4±0.0	13.7±0.0	32.5±0.0	23.3±0.0	19.5
ZS-CLIP (PC) [50]	–	✓	9.9±0.0	28.3±0.0	32.9±0.0	24.0±0.0	23.8
PromptStyler	–	–	**13.8**±1.7	**39.8**±1.3	**38.0**±0.4	**30.3**±0.3	**30.5**

Thanks