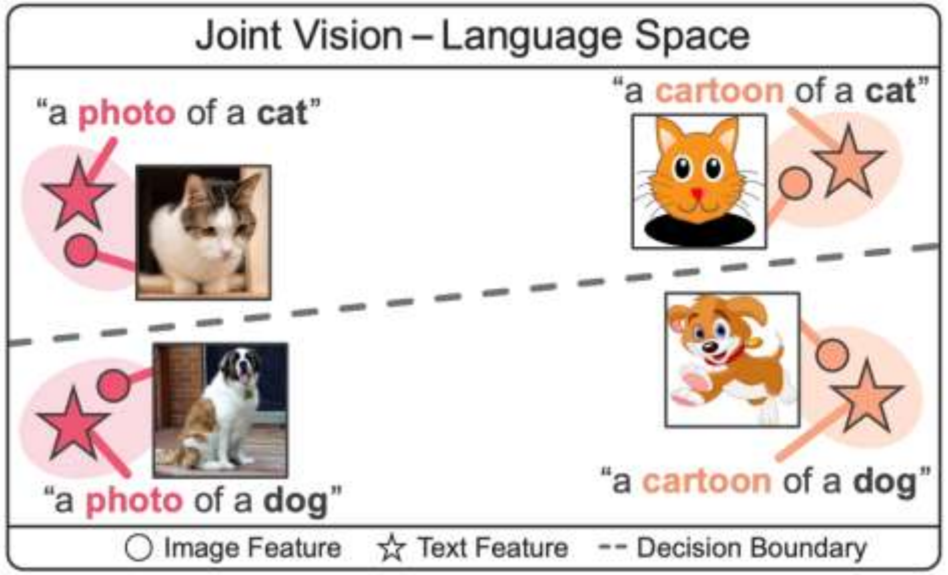
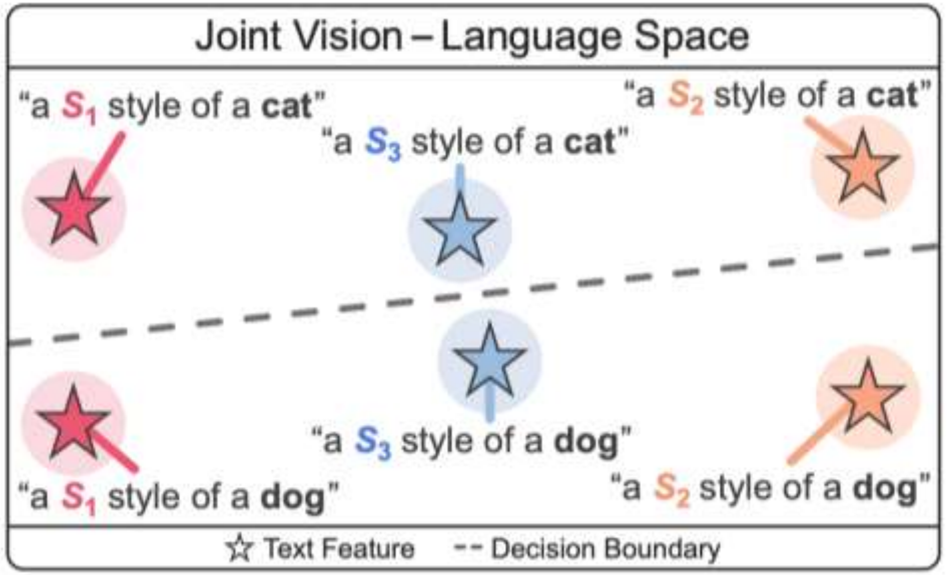
PromptStyler:
Prompt-driven Style Generation
for Source-free Domain Generalization

Represent various image styles in a joint vision language space exploit text features

- Synthesize a variety of styles in a joint vision-language space via prompts to effectively tackle source-free domain generalization

***Style word vector is $K = 80$**

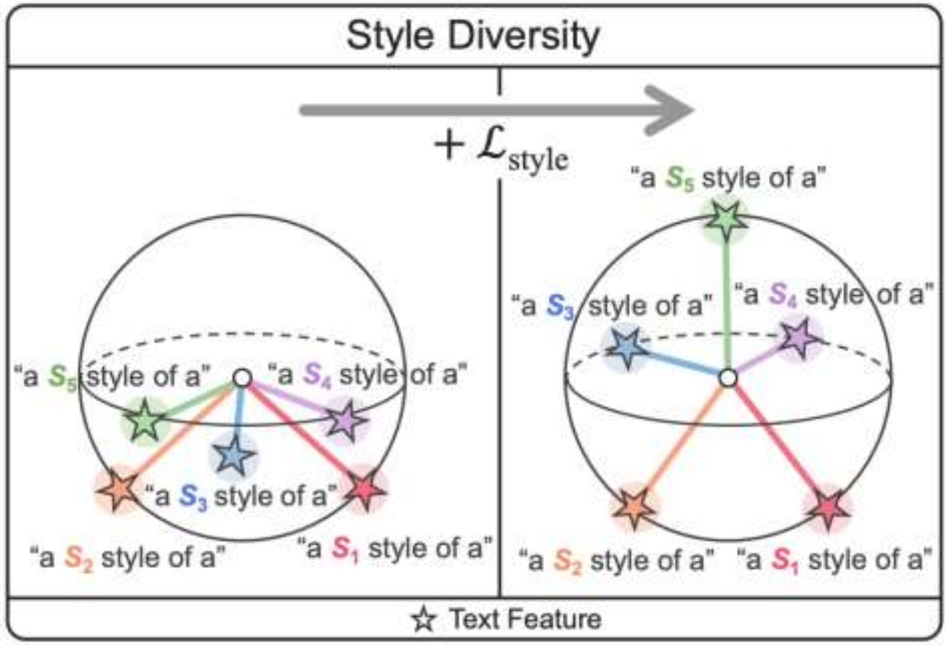




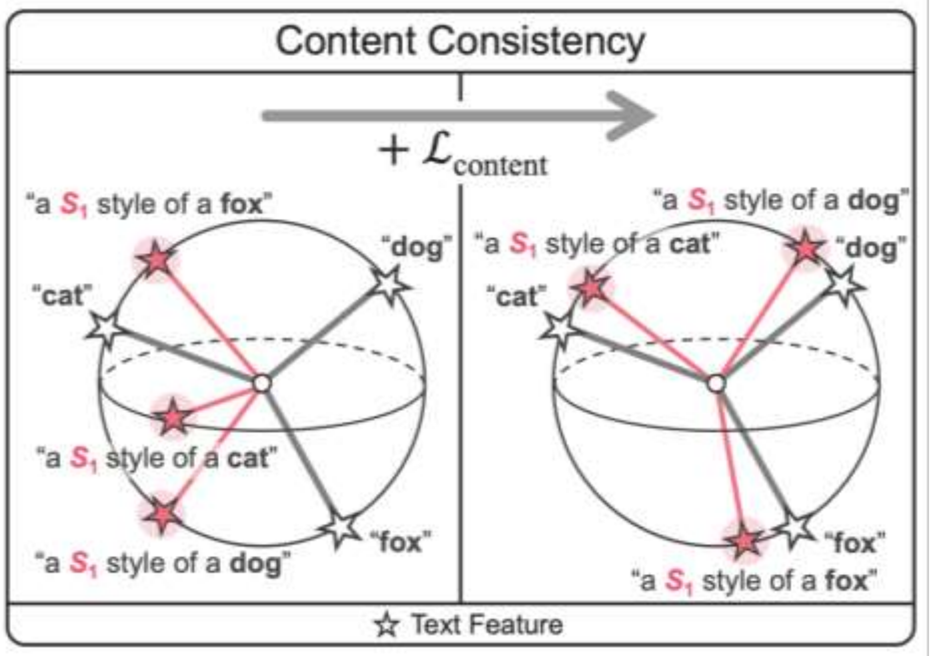
Cho, Junhyeong, et al. "Promptstyler: Prompt-driven style generation for source-free domain generalization." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Prompt-driven style generation

- Learned style word vectors are used to synthesize style content features for training a classifier; these **synthesized features could simulate images of known contents with diverse unknown styles in the joint space**



+


Content Consistency
$+\mathcal{L}_{\text{content}}$
"a S_1 style of a **fox**"
"cat"
"dog"
"a S_1 style of a **cat**"
"a S_1 style of a **dog**"
"fox"
"a S_1 style of a **dog**"
"a S_1 style of a **cat**"
"dog"
"cat"
"fox"
"a S_1 style of a **fox**"
☆ Text Feature

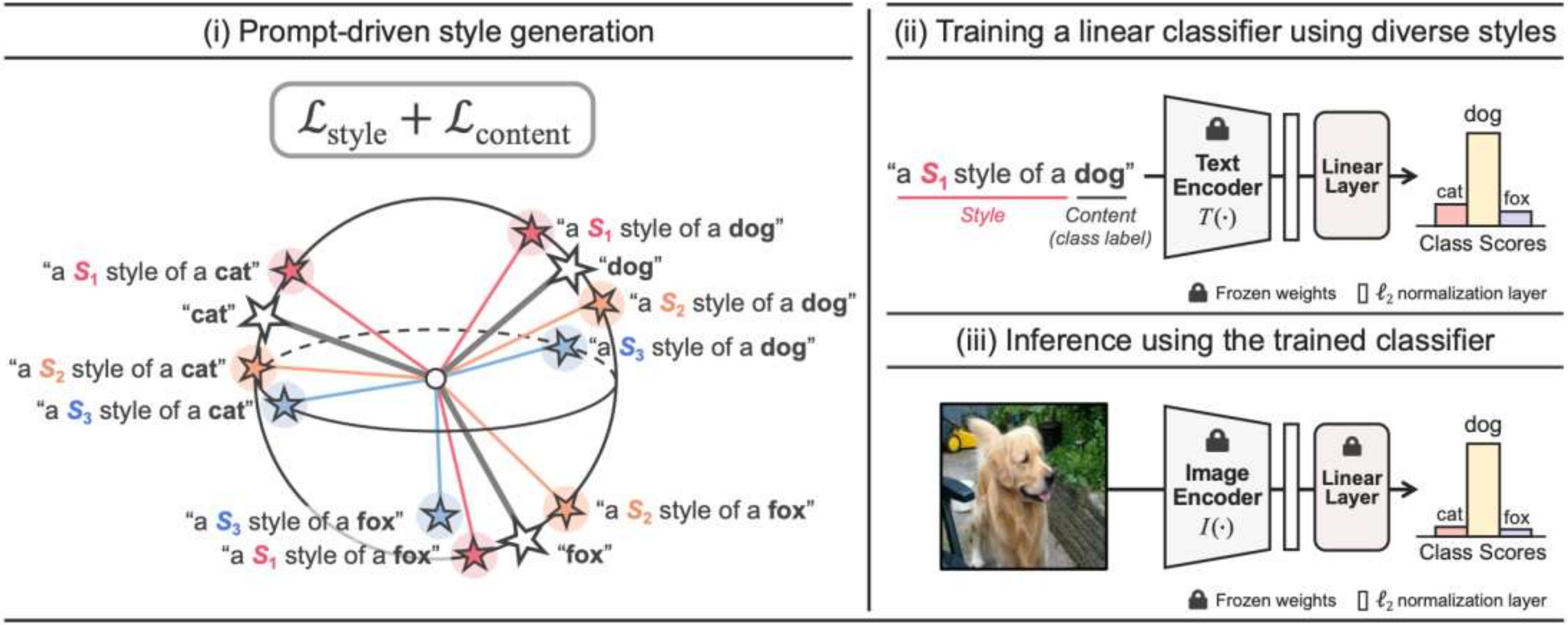
$$\mathcal{L}_{\text{style}} = \frac{1}{i-1}\sum_{j=1}^{i-1}\left|\frac{T(\mathcal{P}_i^{\text{style}})}{\|T(\mathcal{P}_i^{\text{style}})\|_2}\cdot\frac{T(\mathcal{P}_j^{\text{style}})}{\|T(\mathcal{P}_j^{\text{style}})\|_2}\right|$$

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

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

Training a linear classifier using diverse styles

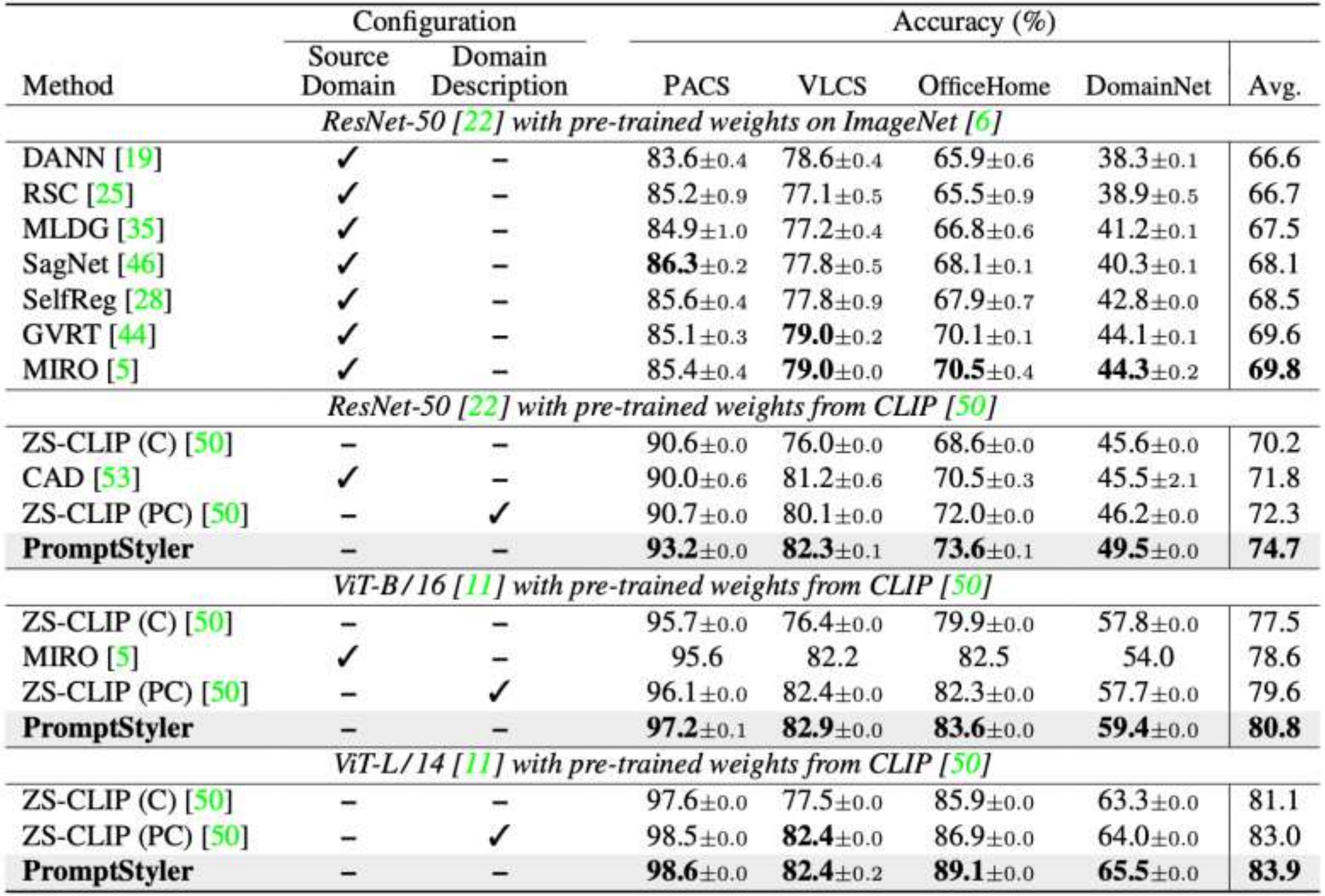
***ArcFace:** $W_j^T x_i = \|W_j\| \, \|x_i\| \cos \theta_j,$



Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

Results

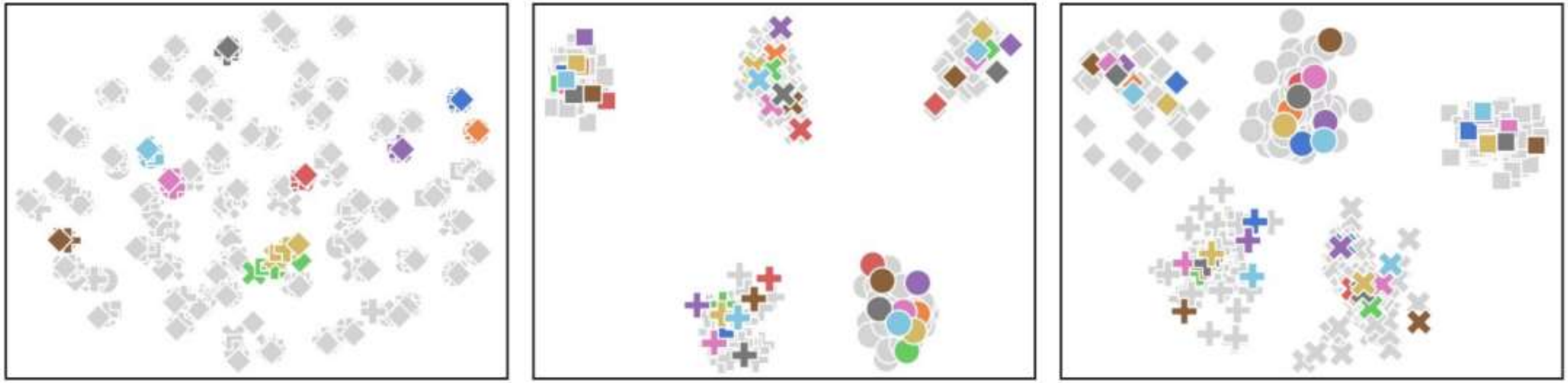
- Comparison with the state-of-the-art domain generalization methods.



	Configuration		Accuracy (%)				
Method	Source Domain	Domain Description	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**

t-SNE visualization results

- t-SNE visualization results for the target task VLCS (5 classes) using synthesized style-content features
  - 5 classes, 80 style word vectors



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

QnA

zaqxsw0526@gmail.com