Table 1: Comparison of our proposed method with the State-of-the-art methods on the DomainBed benchmark with ViT-B/16 backbone. denotes frozen CLIP ViT-B/16 encoder; denotes fine-tuning the entire CLIP ViT-B/16 encoder, **Red** and indicate the best performance in each group.

Model	Venue	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg
CLIP Zero-Shot	-	96.2	81.7	82.4	33.4	57.5	70.2
CoOp	IJCV'22	96.0	81.1	83.5	47.0	59.8	73.5
CoCoOp	CVPR'22	95.7	83.1	84.3	50.4	60.0	74.7
StyLIP [R3]	WACV'24	**98.1**	86.9	84.6	-	62.0	-
MaPLE [R6]	CVPR'23	97.6	85.1	83.4	-	60.4	-
CLIPCEIL	Ours	97.6 ± 0.1	$\mathbf{88.4 \pm 0.4}$	$\mathbf{85.4 \pm 0.2}$	$\mathbf{53.0 \pm 0.3}$	$\mathbf{62.0 \pm 0.1}$	$\mathbf{77.3 \pm 0.2}$
MIRO	ECCV'22	95.6	82.2	82.5	54.3	54.0	73.7
VLV2-SD [R7]	CVPR2024	96.7	83.3	87.4	58.5	62.8	77.7
CLIPCEIL++	Ours	97.2 ± 0.1	85.2 ± 0.5	$\mathbf{87.7 \pm 0.3}$	62.0 ± 0.5	$\mathbf{63.6 \pm 0.2}$	$\mathbf{79.1 \pm 0.2}$

Table 2: Comparison with SAGM and DomainDrop with ResNet-50 backbone on OfficeHome dataset

Model	A	C	P	R	Avg
SAGM	–	–	–	–	70.1
SWAD [R2]	66.1	57.7	78.4	80.2	70.6
DomainDrop	67.3	60.4	79.1	80.2	71.8
DISPEL [R9]	71.3	59.4	80.3	82.1	73.3
CLIP Zero-shot	74.6	49.5	79.4	83.5	71.8
CLIPCEIL	76.9	54.3	85.0	86.3	75.6

Table 3: Performance comparison with text encoder adapter with ViT-B/16 backbone.

Model	A	C	P	R	Avg
Visual + text multi-scale adapter	85.7	70.5	92.0	91.8	85.0
CLIPCEIL (Only visual multi-scale adapter)	86.0	71.2	92.2	92.3	85.4

Table 4: Performance of a linear layer adapter g on OfficeHome dataset with ViT-B/16 backbone

Model	A	C	P	R	Avg
CLIP Zero-shot	82.7	68.0	88.3	90.7	82.4
One linear projector	84.0	69.8	90.2	90.8	83.7
One linear projector $+\mathcal{L}_{\text{ref}}+\mathcal{L}_{\text{dir}}$	85.0	70.6	91.7	91.8	84.8
Average-pooling	84.2	68.6	90.8	91.3	83.7
Two-layer MLP	85.5	70.2	90.7	91.6	84.5
CLIPCEIL (w/ Transformer layer)	**86.0**	**71.2**	**92.2**	**92.3**	**85.4**

Table 5: Accuracy on ImageNet with various domain shifts on ViT-B/16 backbone.

Model	ImageNet	V2	S	A	R	Avg
CLIP Zero-Shot	66.7	60.8	46.1	47.8	74.0	57.2
CoOp	71.5	64.2	48.0	49.7	75.2	59.3
CLIPCEIL	71.6	64.6	49.2	50.5	76.8	60.3

Table 6: Performance with different ViT backbone.

Model	A	C	P	R	Avg
CLIP (ViT-L/14) Zero-shot	89.8	74.8	93.6	94.1	88.1
CLIPCEIL (ViT-L/14)	91.1	79.6	94.8	95.1	90.2
CLIP (ViT-B/32) Zero-shot	82.7	61.8	86.6	88.6	79.9
CLIPCEIL (ViT-B/32)	84.2	66.4	90.0	91.5	83.0

Table 7: Ablation study of contrastive loss on ViT-B/16 backboone.

Model	A	C	P	R	Avg
CLIP+CEIL (w/o $\mathcal{L}_{\text{dir}}$)	83.5	70.0	91.3	90.7	84.1
CLIP+CEIL (w $\mathcal{L}_{\text{dir}}$)	86.0	71.2	92.2	92.3	85.4
SLIP+CEIL (w/o $\mathcal{L}_{\text{dir}}$)	85.4	71.0	91.5	91.3	84.8
SLIP+CEIL (w $\mathcal{L}_{\text{dir}}$)	86.2	72.7	92.2	92.2	85.8

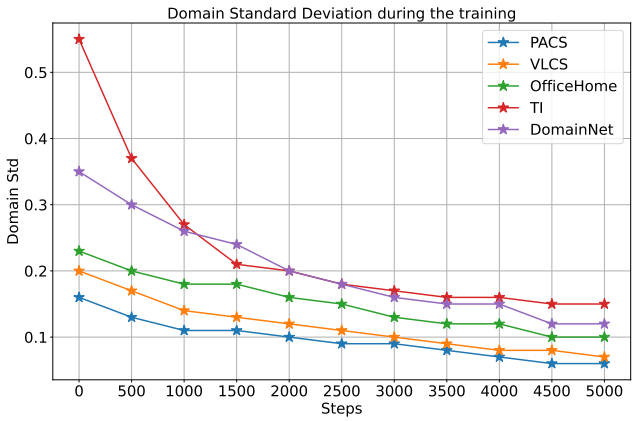
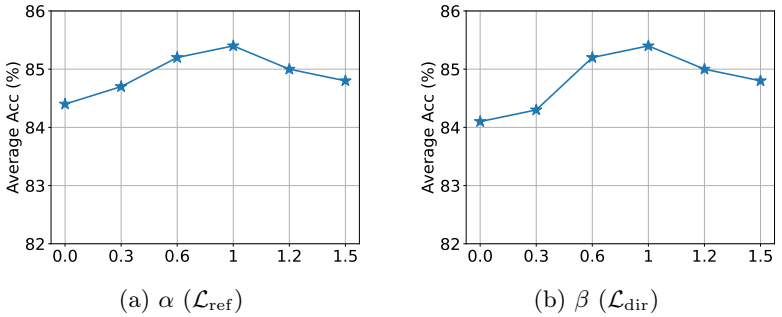


Figure 1: Domain Standard Deviation during the training.



(a) α ($\mathcal{L}_{\text{ref}}$) (b) β ($\mathcal{L}_{\text{dir}}$)

Figure 2: Performance of CLIPCEIL with different hyperparameter α and β on OfficeHome dataset.