CLIPCEIL: Domain Generalization through CLIP via Channel rEfinement and Image-text aLignment

Xi Yu, Shinjae Yoo, Yuewei Lin

Artificial Intelligence Department, Computing and Data Science Directorate,

Brookhaven National Laboratory, Upton, NY 11973.

@BrookhavenLab

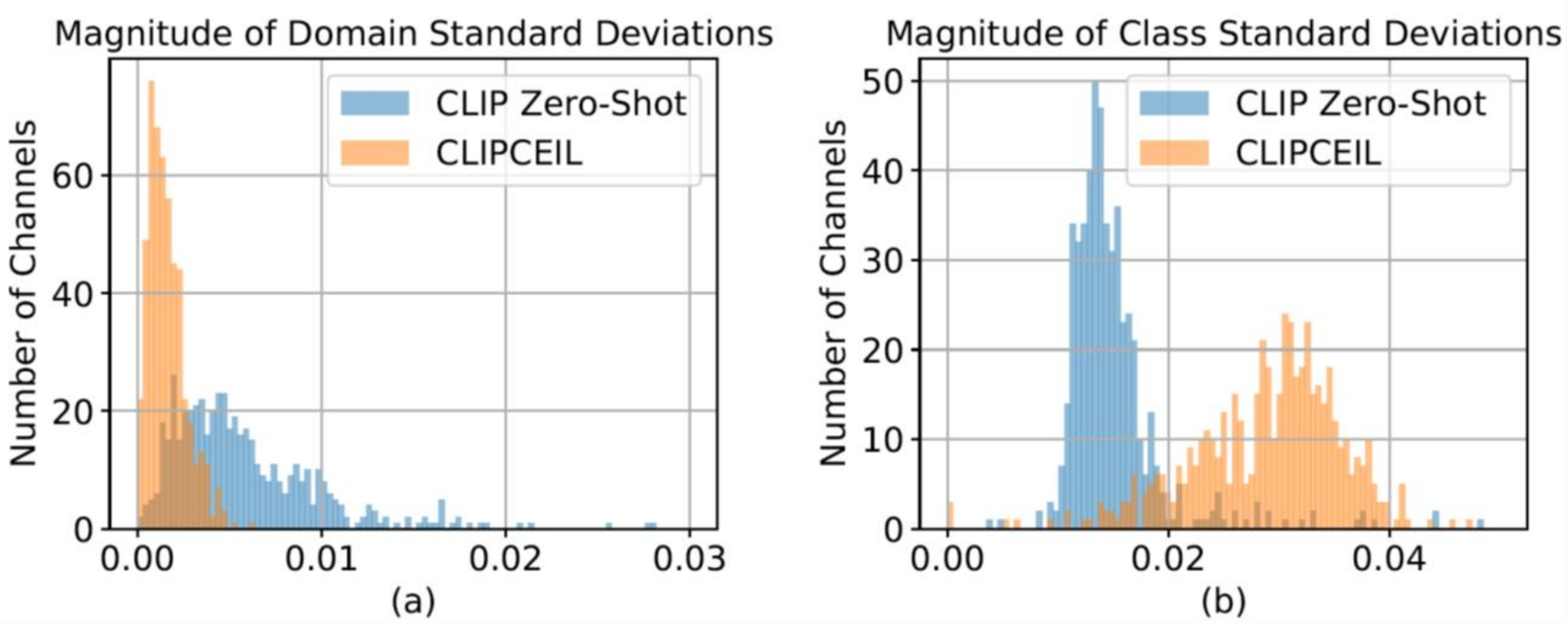
Problem Statement and Contributions

Problem Statement: Domain generalization (DG) addresses the challenge of training a model on one or more distinct but related domains to enable it to generalize effectively to test domains with domain shifts.

Contributions:

- We propose to adapt CLIP through Channel rEfinement and Image-text aLignment (**CLIPCEIL**), ensuring the visual feature channels contain the domain-invariant and class-relevant information while preserving the image-text alignment.
- Our model integrates multi-scale CLIP features by using a self-attention mechanism, technically implemented through one Transformer layer.
- We comprehensively evaluate our proposed method on five widely used Domain Generalization benchmarks. The results demonstrate that our method achieves state-of-the-art performance.

Motivations



Observation: As shown in above Figures (a), many CLIP visual feature channels exhibit unstable activations across domains (illustrated by the blue histogram), indicating a lack of domain invariance. Similarly, as shown in Figure (b), many CLIP visual feature channels show insensitivity, and thus indiscriminative to class variations.

Motivations

Can we enhance the pre-trained model's generalizability by excluding domain-specific (sensitive) and class-irrelevant (insensitive) features?

Model	A	C	P	R	Avg
CLIP full features	82.7	68.0	88.3	90.7	82.4
Channel-Selection	**84.9**	**68.3**	**89.4**	**91.2**	**83.5**

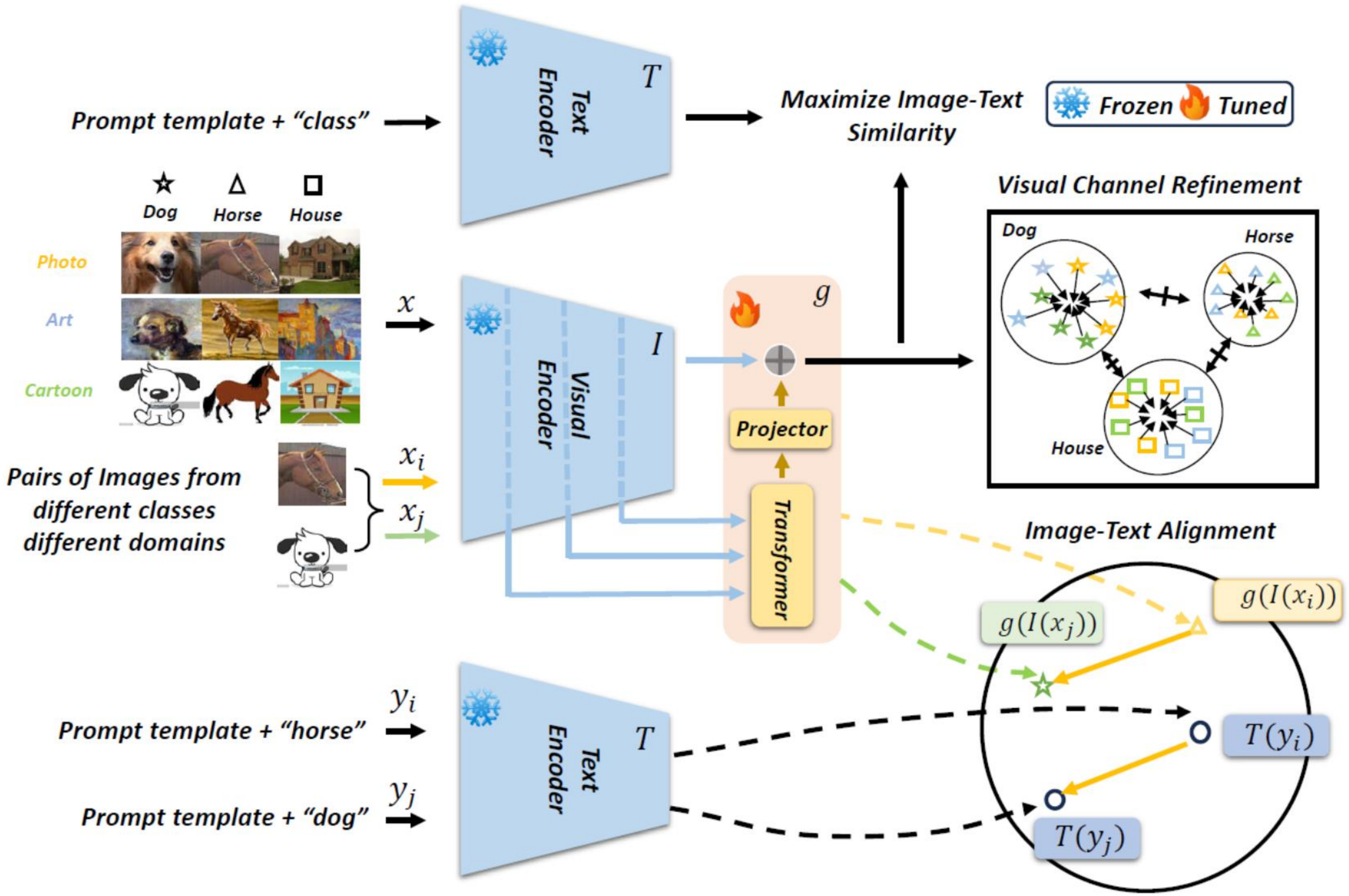
Table 1. Comparison of channel selection (Q=400) with CLIP zero-shot on Office Home benchmark.

To answer it, we conduct a simple experiment using the pre-trained CLIP model on OfficeHome dataset. Given the original 512 CLIP visual feature channels, we select the ones with low domain variance and high class variance. **As shown in Table 1, the simple feature channel selection improves the CLIP zero-shot generalizability.**

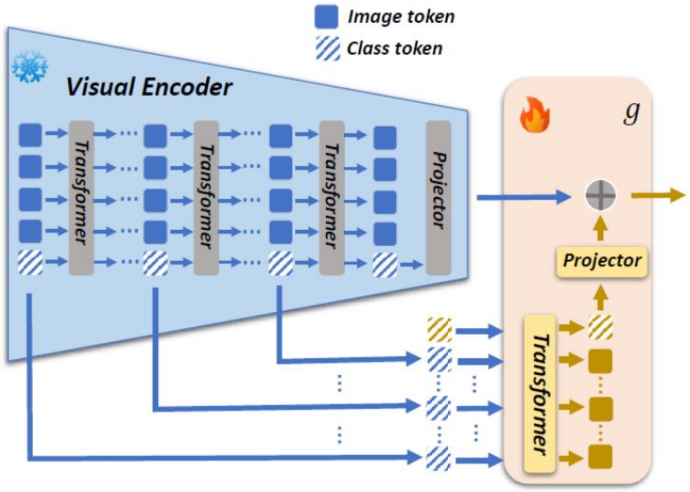
Brookhaven National Laboratory

NEURAL INFORMATION PROCESSING SYSTEMS

Framework Overview



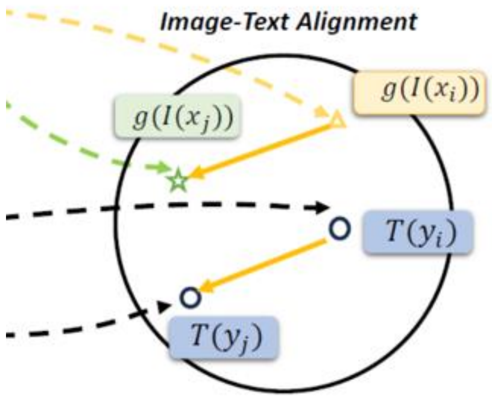
Overview of CLIPCEIL



Architecture of Adapter g

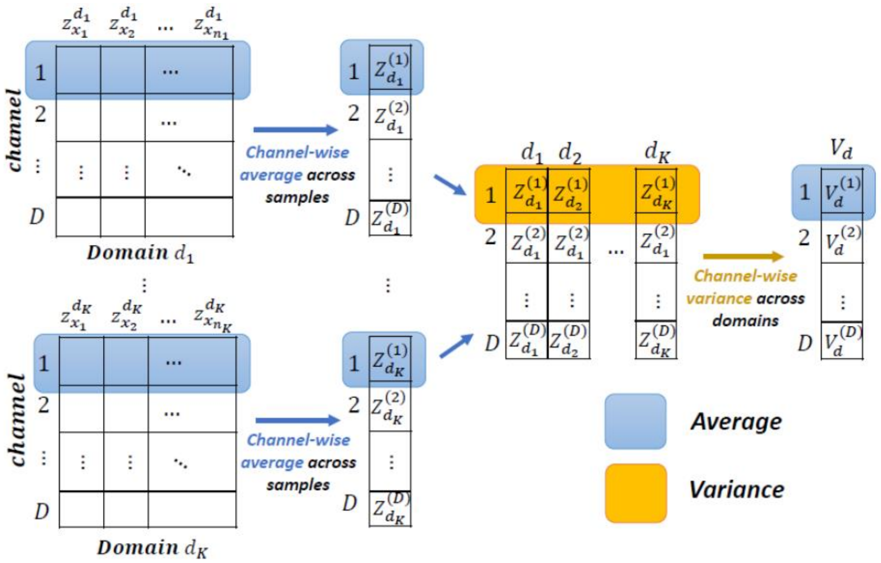
Methodology

Image-Text Alignment Loss:



$$\mathcal{L}_{\mathrm{CE}} = \text{Cross-entropy}\big(\text{Softmax}[g_\theta(I(\mathbf{x})) \cdot \mathbf{T}_y], y\big)$$

$$\mathcal{L}_{\mathrm{dir}} = 1 - \left(\frac{g_\theta(I(\mathbf{x}_i)) - g_\theta(I(\mathbf{x}_j))}{\|g_\theta(I(\mathbf{x}_i)) - g_\theta(I(\mathbf{x}_j))\|} \cdot \frac{\mathbf{T}_{y_i} - \mathbf{T}_{y_j}}{\|\mathbf{T}_{y_i} - \mathbf{T}_{y_j}\|} \right)$$



Channel Refinement Loss :

$$\mathcal{L}_{\mathrm{ref}} = \frac{1}{D} \sum_{m=1}^{D} \log\left(1 + \frac{\sqrt{V_d^{(m)}}}{\sqrt{V_c^{(m)}}}\right)$$

Overall Objective:

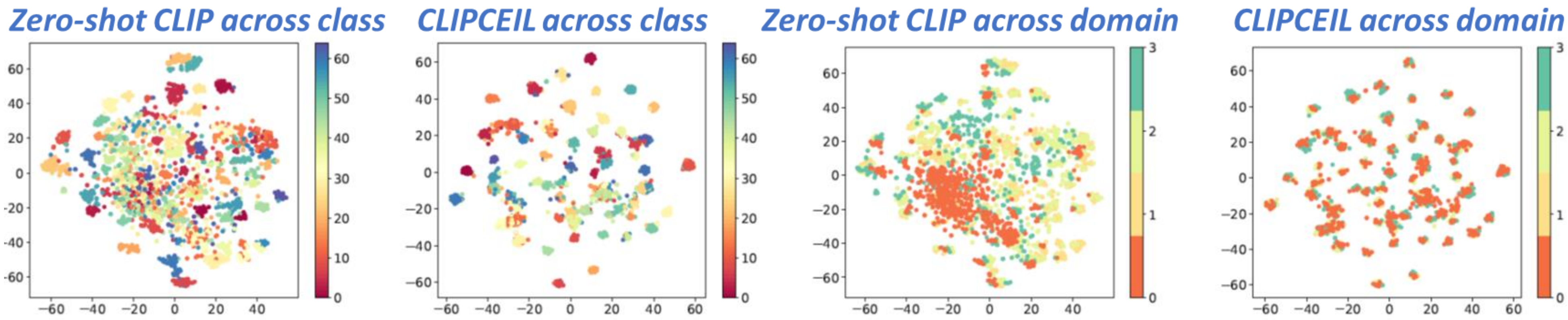
$$\min_\theta \mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{ref}} + \mathcal{L}_{\mathrm{dir}}$$

Results

■ denotes ResNet-50 backbone; ■ denotes frozen CLIP ViT-B/16 encoder; ■ denotes fine-tuning the entire CLIP ViT-B/16 encoder, * denotes the two rounds inference-time fine-tuning.
Red and ■ indicate the best performance in each group.

Comparison with the State-of-the-art methods

Model	Venue	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg
SAGM [54]	CVPR'23	86.6	80.0	70.1	48.8	45.0	66.1
DomainDrop [17]	ICCV'23	89.5	78.3	71.8	-	44.4	-
CLIP Zero-Shot	-	96.2	81.7	82.4	33.4	57.5	70.2
Lin.Probing	-	96.5	82.6	80.4	50.2	57.6	73.5
CoOp [68]	IJCV'22	96.0	81.1	83.5	47.0	59.8	73.5
CoCoOp [67]	CVPR'22	95.7	83.1	84.3	50.4	60.0	74.7
CLIP-Adapter [15]	IJCV'24	96.4	84.3	82.2	-	59.9	—
MaPLE [24]	CVPR'23	97.6	85.1	83.4	-	60.4	-
DPL [62]	2023	97.3	84.3	84.2	52.6	56.7	75.0
StyLIP [4]	WACV'24	98.1	86.9	84.6	-	62.0	-
CLIPCEIL	Ours	97.6 ± 0.1	88.4 ± 0.4	85.4 ± 0.2	53.0 ± 0.3	62.0 ± 0.1	77.3 ± 0.2
MIRO [7]	ECCV'22	95.6	82.2	82.5	54.3	54.0	73.7
CLIPood [44]	ICML'23	**97.3**	85.0	87.0	60.4	63.5	78.6
CAR-FT [35]	IJCV'24	96.8	85.5	85.7	61.9	62.5	78.5
UniDG* [63]	arXiv'23	96.7	**86.3**	86.2	**62.4**	61.3	78.6
VLV2-SD [1]	CVPR'24	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	$\mathbf{63.6 \pm 0.2}$	$\mathbf{79.1 \pm 0.2}$



t-SNE visualization on image features