Unknown Prompt, the only Lacuna: Unveiling CLIP's Potential for Open Domain Generalization

Mainak Singha[1], Ankit Jha[2], Shirsha Bose[3], Ashwin Nair[4], Moloud Abdar[5], Biplab Banerjee[2]

[1]Aisin Corporation, Japan, [2]IIT Bombay, India, [3]TU Munich, Germany, [4]IISER Thiruvananthapuram, [5]Deakin University, Australia

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivation:

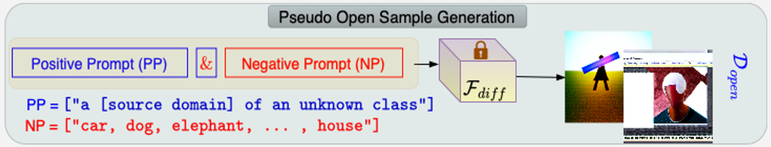
- CNN based generalized models can effectively learn distribution shifts, but often misclassify novel classes during inference, especially when these classes were absent during training.
- Even zero-shot transfer in CLIP and generalized transfer in existing vision language models (VLMs) encounter challenges in detecting completely unlabeled outlier or **unknown** classes, as they rely on class names in the textual input prompt.

Our Proposal:

- Leveraging the pre-trained stable diffusion model for the generation of the pseudo-open training samples (**unknown**) using positive and negative prompts.
- Introducing a novel domain-agnostic prompt learning strategy to generalize the domain-specific style information in the text space.
- Enhance visual semantics in a learnable latent space by maximizing the similarities between the CLIP's textual and visual features.

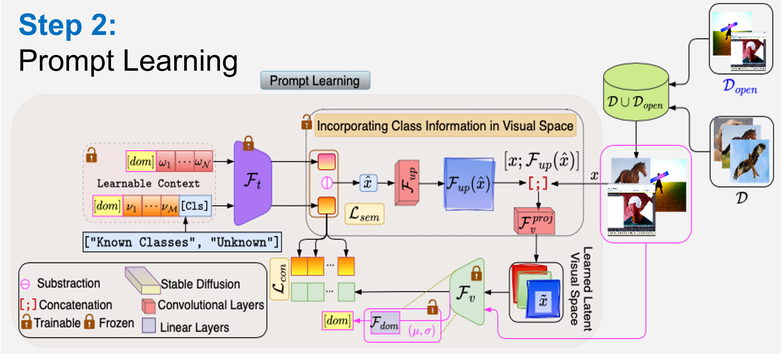
Step 1:

Pseudo Open Sample Generation



Step 2:

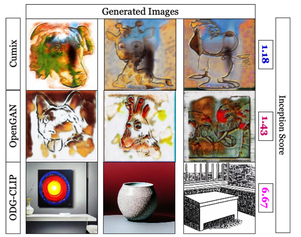
Prompt Learning



Proposed Methodology:

- Aim to improve the quality and semantic versatility of pseudo-open samples by leveraging the pretrained Stable Diffusion v1-5 model. This entails utilizing **unknown** or unseen categories as positive prompts, while known classes serve as negative prompts (indicating what we do not want to generate).
- Our classification strategy encompasses C+1 classes, where the (C+1)-th index is designated for the novel unknown-class.
- We advocate for adaptive prompt learning across all classes, enabling the capture of domain-specific distributions and overarching semantic contents through distinct token sets.
- We refine visual-textual contrastive learning for ODG-CLIP by enhancing the discriminability of visual embeddings. This is achieved by establishing a latent visual space, guided by the prompts we've developed.
- Also, we propose a cross-domain semantic consistency loss to cultivate a robust class-wise correlation in the derived latent visual representations across images from different domains but sharing identical class labels.

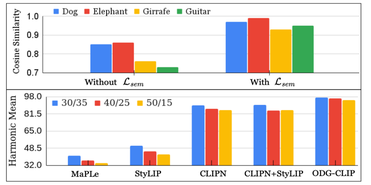
Experiments and Results:



Methods	PACS		VLCS		OfficeHome		Digits-DG		Multi-Dataset		Mini-DomainNet		Average	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CNN-based														
Cumix [33]	57.85	41.05	52.46	50.11	51.67	49.40	58.13	54.20	42.18	46.91	50.27	39.16	52.09	46.81
MixStyle [68]	63.35	48.30	52.30	50.61	53.52	49.53	60.23	56.35	42.18	46.91	50.43	40.25	53.67	48.66
DAML [48]	65.49	51.88	53.53	51.59	56.45	53.34	59.51	55.61	46.61	51.71	52.81	43.63	55.73	51.29
MEDIC [27]	89.81	83.03	57.28	55.73	60.26	57.91	83.28	66.30	50.74	53.13	55.29	45.71	66.11	60.30
CLIP-based														
CLIP [41]	95.16	76.77	91.84	72.94	81.43	63.62	77.08	61.95	77.88	72.19	84.50	68.94	84.65	69.40
CLIP + OpenMax [3]	93.45	79.13	92.09	73.67	81.00	61.54	76.93	62.78	78.34	73.26	81.89	69.40	83.95	69.96
CLIP + OSDA [31]	92.62	75.40	90.21	70.89	82.58	67.35	80.53	65.70	74.45	75.22	82.00	73.62	83.73	71.36
CoOp [67]	78.77	26.87	92.02	39.26	73.85	36.26	58.54	34.81	66.03	44.34	61.13	68.34	71.72	41.65
CoCoOp [65]	85.76	32.93	89.47	37.01	75.38	34.38	52.77	33.50	64.84	47.57	60.63	56.30	71.48	40.28
MaPLe [29]	93.97	48.47	89.70	43.33	79.47	33.06	70.54	43.83	69.34	62.20	74.67	60.57	79.62	48.58
LASP [1]	88.45	30.37	90.67	39.41	76.13	34.52	60.89	35.23	66.78	50.22	62.34	61.56	74.21	41.89
PromptSRC [21]	94.53	43.32	90.13	42.78	80.21	36.40	75.34	44.25	65.51	59.45	73.60	62.56	79.89	48.13
CLIPN [55]	96.24	45.00	84.82	50.72	84.55	42.83	81.70	45.56	77.16	62.60	77.38	66.92	83.64	52.27
StyLIP [1]	95.36	50.74	90.75	65.66	84.73	60.97	80.59	58.15	79.88	71.99	80.22	69.11	85.26	62.77
CLIPN + StyLIP	96.37	64.46	84.65	68.02	83.67	76.50	82.14	59.24	76.93	72.15	86.59	76.18	85.06	69.43
MaPLe + SD	91.47	82.60	91.70	72.67	85.02	80.60	79.92	65.82	77.62	72.83	83.79	79.30	84.92	75.64
PromptSRC + SD	93.21	87.95	90.34	72.62	84.60	83.31	80.92	65.37	78.44	77.89	83.87	82.95	85.23	78.35
StyLIP + SD	91.78	87.42	92.11	73.34	85.51	81.22	81.45	68.10	79.05	78.52	84.12	83.21	85.67	78.64
ODG-CLIP	**99.53**	**99.70**	**95.71**	**86.53**	**98.32**	**96.08**	**91.53**	**78.27**	**84.60**	**90.00**	**95.68**	**94.48**	**94.23±0.19**	**90.84±0.26**

Methods	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
Only PP	92.45	92.16	93.72	90.83	78.20	81.57	91.30	89.78
PP + NP	**99.53**	**99.70**	**98.32**	**96.08**	**84.60**	**90.00**	**95.68**	**94.48**

Methods	PACS		O.H.		M.Data		M.DNet	
	Acc	H	Acc	H	Acc	H	Acc	H
w/o $\hat{x}$ and $\mathcal{L}_{sem}$	90.47	88.34	92.21	87.00	73.56	75.73	87.24	83.51
w/o $\mathcal{L}_{sem}$, with $\hat{x}$	94.21	92.56	95.67	91.56	80.34	85.32	91.24	90.88
Manual $\hat{x}$	93.54	92.82	95.31	91.22	78.53	79.26	90.65	86.52
Full (ours)	**99.53**	**99.70**	**98.32**	**96.08**	**84.60**	**90.00**	**95.68**	**94.48**



Summary:

- Open Domain Generalization (ODG) addresses both known and novel classes across out-of-distribution domains. During inference, the target domain may include **completely unlabeled** familiar categories and entirely **new categories**, resembling real-world recognition scenarios.
- Our proposed ODG-CLIP harness the semantic capabilities of the vision-language pretrained model CLIP, to solve the ODG task, just using a text prompt of **unknown** class.