DPA : Dual Prototypes Alignment for Unsupervised Adaptation of Vision-Language Models

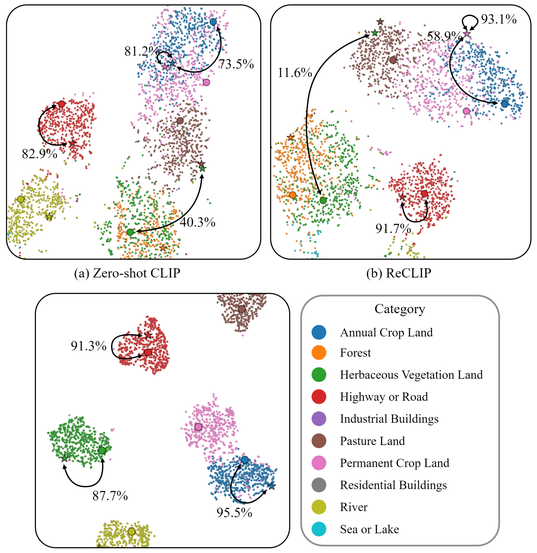
Eman Ali [1] Sathira Silva [1] Muhammad Haris Khan [1]

[1]Mohamed Bin Zayed University of Artificial Intelligence

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Background and Problem Statement

How can we adapt CLIP to target domains using unlabeled data while addressing challenges such as noisy pseudo-labels and modality gaps?



Limitations of Existing Approaches

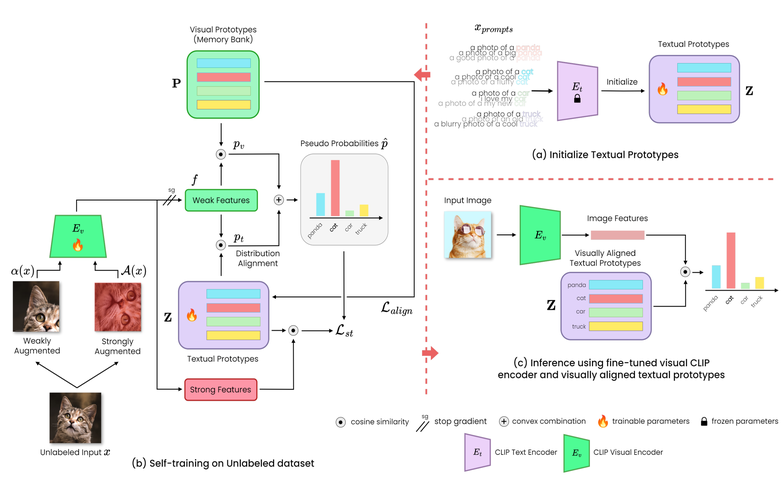
- Pseudo-labeling faces challenges due to noisy labels and modality gaps.
- Projection spaces and label propagation become costly and inefficient for datasets with a large number of classes.

Contributions

- Proposed a novel dual-modality prototypic alignment framework for the unsupervised domain adaptation of vision-language models (VLMs).
- Introduced ranking-based pseudo-labels to mitigate noise.

DPA Architecture

- **Dual-Modality Prototypes:** Image prototypes serve as non-parametric classifiers, reducing noise, while textual prototypes initialized using zero-shot CLIP enhance semantic alignment.
- **Convex Combination for Pseudo-Labels:** Combines the outputs of image and textual prototypes for accurate pseudo-labeling.
- **Noise Mitigation:** Ranks pseudo-labels in the classification loss to mitigate noise, particularly during early training stages.
- **Alignment of Visual and Textual Prototypes:** Bridges modality gaps by aligning textual prototypes with image prototypes.



Experiments – Datasets and Evaluation Metrics

- **Datasets Methods:** Evaluated on 13 diverse datasets, including general classification, specialized domains, fine-grained tasks, and more.
- **SOTA Methods:** Zero-shot CLIP; **Unsupervised Adaptation Methods for CLIP:** UPL, POUF, and LaFTer.
- **Adaptation:** Fine-tuned layer normalization of the image encoder and textual prototypes.

Experiments: Comparative Results

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFTer	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

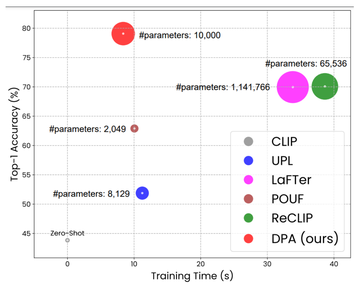
Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

Experiments: Component Analysis

Method	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	82.40	66.46	87.50	58.74	89.80	65.10	64.20	64.79
Base	93.57	48.99	61.94	19.20	84.10	68.45	90.24	59.25	95.95	73.55	65.45	69.15
Center	95.44	55.53	70.56	19.80	84.65	75.27	**90.71**	61.53	95.96	76.01	67.30	72.07
Center+w	95.46	54.54	**80.06**	19.56	84.63	75.44	90.49	61.19	**95.97**	75.92	67.51	72.80
Center+w+Align (DPA)	**96.06**	**55.69**	80.04	**20.67**	**84.76**	**75.56**	**90.71**	**62.62**	**95.97**	**76.47**	**68.49**	**73.37**

Experiments: Efficiency Comparison

DPA outperforms the baselines in terms of training efficiency and performance, requiring fewer parameters and reducing computational complexity.



Conclusion

- DPA bridges the domain gap between visual and textual representations in VLMs.
- Introduces dual prototypes as classifiers and ranks pseudo-labels for robust self-training.
- Outperforms zero-shot CLIP and state-of-the-art methods across 13 downstream tasks.