OpenDAS: Domain Adaptation for Open-Vocabulary Segmentation

Gonca Yilmaz[1,2*] Songyou Peng[1] Francis Engelmann[1] Marc Pollefeys[1,3] Hermann Blum[1]
[1] ETH Zurich [2] University of Zurich [3] Microsoft

Abstract

The advent of visual language models (VLMs) has transformed image understanding from fixed classifications to dynamic image-language interactions. Despite their flexibility, VLMs often trail behind closed-set classifiers in accuracy, due to their heavy dataset demands and absence of domain-specific knowledge. Hence, we introduce a new task domain adaptation for open-vocabulary segmentation. We propose a novel approach that leverages prompt tuning in combination with triplet loss training to bypass extensive fine-tuning. Our work demonstrates improved performance and practicality for real-world applications, achieving +3.9% accuracy in ScanNet++ office subset with unseen classes during training compared to other parameter-efficient learning strategies and on 2D semantic segmentation benchmarks +5.8% on KITTI-360 and +5.4% on ADE20K-150 validation split. We further demonstrate the benefit of semantic understanding in open-vocabulary 2D semantic segmentation by replacing CLIP in an existing pipeline, improving mIoU by 6.0% on ADE20K-150 validation split, and in open-vocabulary 3D instance segmentation improving the AP score by 4.1% on ScanNet++ Offices.

1. Introduction

Visual-language models (VLMs) like CLIP [11] and ALIGN [4] have transformed image understanding, enabling open-vocabulary segmentation for precise object localization with language cues. Despite their broad applications — from robots understanding textual commands to text-based navigation systems—these models struggle with specialized domains and fine-grained segmentation, primarily due to their reliance on extensive datasets for learning.

Addressing these limitations, we introduce a new task, *domain adaptation for open-vocabulary segmentation*, focusing on enhancing CLIP-based models' performance through prompt tuning. This novel strategy allows for parameter-efficient adaptation to target domains, improving segmentation by learning to match text and image crops independently of mask prediction. By combining cross-entropy and triplet loss, we ensure the model learns new classes while retaining its generalization capabilities.

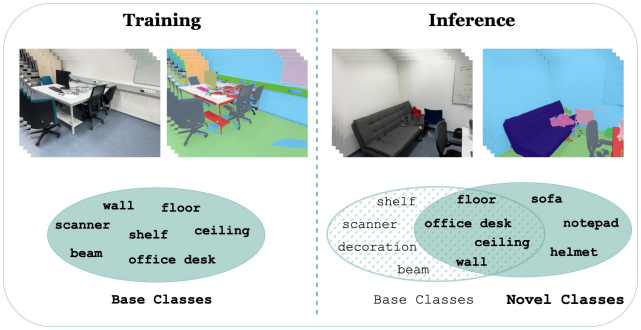

Figure 1. **Domain Adaptation for Open-Vocabulary Segmentation.** We introduce a new task, *domain adaptation for open-vocabulary segmentation*, where we adapt to a specific domain with a set of segments and base classes, but test on visually closer domains with base and novel (unseen) classes, demonstrating generalization capabilities while still adapting to the domain.

Our extensive experiments across various datasets demonstrate superior performance over existing uni-modal and multi-modal adaptation strategies, significantly improving segment classification and semantic understanding, even with predicted masks. Our contributions include a new model architecture rooted in previous multi-modal prompt tuning works and the innovative use of triplet loss for domain adaptation in open-vocabulary segmentation, setting a new benchmark in the field.

2. Related Work

Open-vocabulary semantic segmentation, a method for localizing objects using language cues, has evolved with models like CLIP [11] and ALIGN [4] facilitating comparison between class-agnostic masks and textual embeddings. Approaches like LSeg [8] and OpenSeg [3] have advanced this field by aligning features at different granularity levels with text embeddings for accurate segmentation. Notable meth-

*corresponding author: gyilmaz@student.ethz.ch

ods include leveraging CLIP for both textual and image embeddings in segmentation tasks [2, 3, 8, 9]. Inspired by the advancements in 2D, OpenMask3D [13] also leverages CLIP for 3D instance segmentation.

Domain adaptation for Visual Language Models (VLMs) has focused on improving accuracy through fine-tuning, with methods like WiSE-FT [14]. Prompt learning emerged as a less computationally intensive alternative for adapting models, recently in VLMs [5–7, 18, 19].

Among these notable developments, CoCoOp stands out for adding dynamic textual prompts based on image features, enhancing adaptability [18]. On the other hand, VPT introduces learnable visual prompts alongside a linear probe layer to improve model adaptability, highlighting a static but customizable approach to prompt learning [5]. Further research in this field explores the integration of textual and visual prompts, aiming to increase learning efficiency. These methods vary from coupling interim prompts to leveraging distinct encoders and attention masks for adaptation, demonstrating a range of effectiveness in adapting VLMs to new domains [6, 7].

3. Method

We aim to adapt pre-trained vision-language models, e.g. CLIP [11], into the target domain for open-vocabulary segmentation without losing its generalizability. In this section, we first introduce multi-modal prompt tuning and propose a novel training strategy based on a triplet loss (Sec. 3.1), and discuss how to mine data for the triplet loss (Sec. 3.2).

Multi-modal Prompt Tuning. Prompt tuning enables adapting CLIP to target domains by learning the prompts instead of handcrafting them. The additional tokens, $\mathbf{p}_v^{(j)}$ and $\mathbf{p}_t^{(j)}$, are appended to the visual and textual encoder separately. They provide contextual information on target domains while keeping the model as is. Thus, it only introduces a small fraction of new learnable parameters and optimizes the learning process. For more details, we refer the reader to MaPLe [6]. On the contrary to MaPLe, we omit coupling layers, adopting a simpler architecture.

3.1. Optimization

Here we introduce how we optimize the visual prompts $\mathbf{p}_v^{(j)}$ and text prompts $\mathbf{p}_t^{(j)}$ in each layer. The process commences with visual prompts optimization and shifts to text prompts.
Optimization of Visual Prompts. In each iteration, we randomly sample a batch of 16 image segments with their class names, passing through the CLIP visual and text encoder to obtain their embedding $\mathbf{v}_i$ and $\mathbf{t}_i$ for each segment i. Following VPT [5], we use a cross-entropy loss $\mathcal{L}_{\text{ce}}(\mathbf{v}_i, \mathbf{t}_i)$ to optimize the visual prompts $\mathbf{p}_v^{(j)}$ for.
Optimization of Text Prompts. We then freeze the visual prompts and solely optimize the text prompts $\mathbf{p}_t^{(j)}$ with:

$$\mathcal{L}(\mathbf{v}_i, \mathbf{t}_i^+, \mathbf{t}_i^-) = \mathcal{L}_{\text{ce}}(\mathbf{v}_i, \mathbf{t}_i^+) + \lambda \mathcal{L}_{\text{triplet}}(\mathbf{v}_i, \mathbf{t}_i^+, \mathbf{t}_i^-) \quad (1)$$

where $\mathbf{v}_i, \mathbf{t}_i^+$ and $\mathbf{t}_i^-$ are the visual embedding, true class name, and negative class names for segment i, and $\mathcal{L}_{\text{triplet}}$ is a triplet loss [1] with a margin $\mu = 1.5$. The triplet loss helps the model retain open-vocabulary segmentation capabilities while adapting to a specific domain, inspired by the contrastive objective from CLIP [11]. Note that, we gradually increase the λ in (1) to λ_{max}.

3.2. Triplet Mining

Negative Sample Database. One challenge of triplet loss is to find proper negative samples to form triplets. To address this issue, we instruct GPT-4 to generate 5 closely related class names for each category in the training dataset.
Online Hard Negative Sample Mining. After generating a negative database, we apply an online hard negative mining strategy to find informative triplets with the hardest negative name during training, as defined in [12]. This is pivotal in refining the model's ability to distinguish between closely related but distinct classes.

4. Experiments

We first test our approach's open vocabulary and closed-set classification capabilities on the ScanNet++ Offices, KITTI-360, and ADE20K-150 datasets.

4.1. Experiment Setup

Datasets.
ScanNet++ Offices [15] A dataset with 3D indoor scenes. We take a subset of 14 office scenes with 7989 annotated images and 156 labels for training and 16 office scenes with 11054 images and 233 labels for validation. 108 of these labels are in common.
KITTI-360 [10] An outdoor dataset with 11 sequences collected in Germany, which comprises 37 semantic classes.
ADE20K-150 [16, 17] It includes RGB images from both indoor and outdoor scenes with 2000 images for validation with 150 distinct classes.
Baselines. We compare against state-of-the-art prompt learning methods CoCoOp [18], VPT [5], RPO [7], MaPLe [6], and a robust fine-tuning method WiSE-FT [14].
Metrics. For the adaptation performance with ground-truth masks, we use Accuracy (Acc), Weighted-F1 (W-F1), Base-F1 (B-F1) over base classes, and Novel-F1 (N-F1) over novel classes. As for the experiments with predicted masks, we measure the commonly used metrics mean IoU (mIoU) and the mean Accuracy (mAcc) for 2D and Average Precision (AP) and AP_{25} (AP with at least 25% overlap) for 3D.
Implementation Details. We begin by optimizing visual prompts for 5 epochs, with a warmup epoch at a learning

rate of 10^{-5}. From the second epoch onwards, the learning rate is set to 0.0025 using a cosine scheduler. Following this, we optimize text prompts for another 5 epochs with the same base learning rate. Training is conducted with a batch size of 16 using an SGD optimizer on a single NVIDIA A100 GPU, with a total training time of around 10-15 hours depending on the datasets. We set λ_{min} and λ_{max} to 2 and 5.

4.2. Results

Parameter-efficient Domain Adaptation for Visual-Language Models. We apply CLIP to the task of 2D segment classification, thereby augmenting the efficacy of the overall open-vocabulary segmentation process. The results on segment classification are detailed in Table 1. As shown, all prompt learning techniques improve CLIP's segment classification capabilities in the target domain. However, we see that the baseline multi-modal approaches trained with the cross-entropy loss do not achieve necessarily better results than the visual prompt tuning method. This can be attributed to the greedy nature of cross-entropy loss, impacting CLIP's generalization capabilities.

We also present the qualitative results on all three datasets in Fig. 2. Our method shows clear improvements upon the original CLIP predictions. OpenDAS can distinguish similar classes like "door frame" and "door" due to the triplet loss. Original CLIP suffers from similar classes' proximity as seen by the examples of confusion between "wall outlet" - "wall", and "sidewalk" - "road". There are class matching ambiguities, such as the "table" example in the second row classified as "office desk" by CLIP. In contrast, our method can learn the annotator's intention through visual examples introduced during the adaptation.

Open-Vocabulary Segment Classification. We set up 125 novel (unseen) classes for ScanNet++ Offices to further evaluate open-vocabulary segment classification capabilities. As illustrated in Table 1, the original CLIP is indifferent to base and novel classes. Baseline methods all exhibit noticeable performance boosts over the original CLIP on the W-F1 and B-F1, but only marginally improve upon the N-F1 metric. In contrast, we demonstrate superior performance on all three metrics. We attribute the improvement in novel classes to the visual adaptation, closing the domain gap between CLIP's training images and segments. In the meantime, triplet loss preserves the structured embedding space, indicating that our method can adapt to the target domain while preserving strong open-vocabulary capability.

Comparison Against Robust Fine-Tuning Method WiSE-FT [14]. We compare our approach against a robust fine-tuning method that fine-tunes the entire CLIP text and visual encoder. We show comparisons in Table 2 on KITTI-360 and ADE20K on the closed-vocabulary setting, and ScanNet++ Offices on the open-vocabulary setting. As observed, when compared to fine-tuning the CLIP text encoder, we can achieve significant improvements in all metrics and all three datasets by only training ∼0.1% of CLIP-text encoder parameters. As for the comparison over fine-tuning the visual encoder, we show competitive results with only ∼0.05% of the parameters and significantly improve novel classes in ScanNet++ Offices.

Benefit of Semantic Understanding in Open-Vocabulary Segmentation. Moreover, we apply our prompt tuning method to OVSeg [9] and OpenMask3D [13]'s predicted

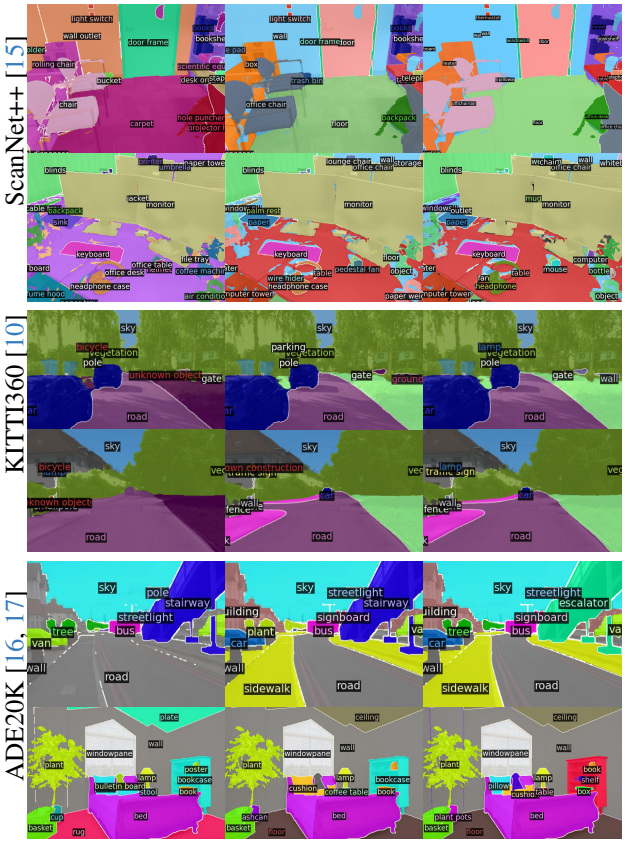

Figure 2. **Qualitative Comparison on Segment Classification.** We show the predicted object classes with the ground truth masks on three datasets. Objects are colorized based on the class id.

Method	Modality	# Params	W-F1	B-F1	N-F1
No adaptation		0	11.2	11.0	12.0
CoCoOp [18]		∼ 77K	25.7	34.3	12.7
VPT [5]		∼ 786K	33.8	37.6	12.8
RPO [7]		∼ 43K	30.6	40.9	14.9
MaPLe [6]		∼ 18935K	36.3	48.1	18.4
OpenDAS (**Ours**)		∼ 233K	**40.2** +3.9	**51.5** +3.4	**23.0** +4.6

Table 1. **Quantitative Comparisons on Open-Vocabulary Segment Classification on ScanNet++ Offices**. We evaluate the ScanNet++ Offices, with weighted-F1 (W-F1) metric and on base and novel classes with base-F1 (B-F1) and novel-F1 (N-F1).

Method	Modality	# Params	ScanNet++ W-F1	ScanNet++ B-F1	ScanNet++ N-F1	KITTI-360 Acc	KITTI-360 W-F1	ADE20K-150 Acc	ADE20K-150 W-F1
No adaptation		0	11.2	11.0	12.0	19.1	23.4	27.8	32.7
WiSE-FT [14]		∼ 123M	31.0	35.0	29.5	53.9	58.8	47.9	51.0
WiSE-FT [14]		∼ 304M	**45.9**	47.3	**45.3**	**78.8**	**80.5**	**73.9**	**74.8**
OpenDAS (**Ours**)		∼ 233K	40.2	**51.5**	23.0	75.7	75.2	73.1	71.9

Table 2. **Comparison with Robust Fine-Tuning [14] on KITTI-360, ADE20K-150, and ScanNet++ Offices.** We outperform the text-encoder fine-tuning setting on all three datasets with about 1000× fewer parameters. We also present competitive results over the image-encoder fine-tuning setting that uses over 2000× more parameters, showing stronger N-F1 results on ScanNet++ Offices.

Method	ADE20K-150 mIoU (%)	ADE20K-150 mAcc (%)	Method	ScanNet++ Offices AP	ScanNet++ Offices AP_{25}
OVSeg [9]	29.8	48.1	OpenMask3D [13]	8.1	14.1
+ OpenDAS	**35.8** +6.0	**51.2** +3.1	+ OpenDAS	**12.2** +4.1	**24.0** +9.9

Table 3. **Benefit of Semantic Understanding in Open-Vocabulary Segmentation.** We apply our prompt tuning method to a recent open-vocabulary 2D semantic segmentation model OVSeg [9] and open-vocabulary 3D instance segmentation model OpenMask3D [13]. Our method helps with understanding semantic labels even with only their predicted masks.

masks, assessing whether our method can help them better understand each segment's semantics. In Table 3, we observe the performance boost in both mIoU and mAcc when applying our method for OVSeg. This demonstrates that OpenDAS can be incorporated into open-vocabulary segmentation pipelines using the predicted masks.

5. Conclusion

Our study shows that prompt tuning combined with triplet loss substantially enhances VLM performance in open-vocabulary segmentation. However, challenges persist, such as base class preference during inference and lack of suitable benchmarks in this emerging field. Future work could integrate the learning process with predicted masks to boost performance, few-shot learning settings and establish custom benchmarks. Nevertheless, our results are promising, making VLMs more adaptable and practical for real-world applications.

References

[1] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *BMVC*, 2016. 2

[2] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2023. 2

[3] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 1, 2

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[5] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2, 3

[6] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 2, 3

[7] D. Lee, S. Song, J. Suh, J. Choi, S. Lee, and H. J. Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, 2023. 2, 3

[8] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 1, 2

[9] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023. 2, 3, 4

[10] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE TPAMI*, 2022. 2, 3

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[13] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, 2023. 2, 3, 4

[14] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 2, 3, 4

[15] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2, 3

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2, 3

[17] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 2, 3

[18] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2, 3

[19] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2