MOVSeg: Open Vocabulary Segmentation from Multi-Modal Inputs

Gonca Yilmaz
University of Zurich
Zurich, Switzerland
gonca.yilmaz@uzh.ch

Songyou Peng
ETH Zurich
Zurich, Switzerland
songyou.peng@inf.ethz.ch

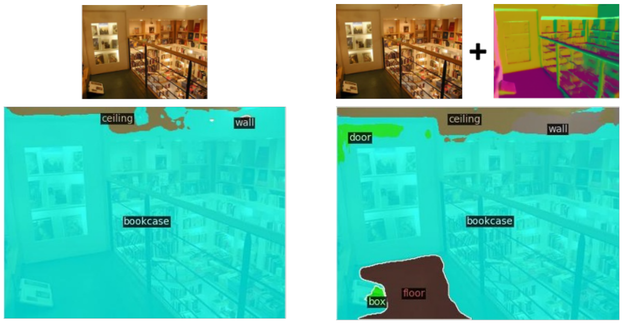
Hermann Blum
ETH Zurich
Zurich, Switzerland
hermann.blum@inf.ethz.ch

Abstract

Trained on large collections of image-text pairs, visual-language models have been shown to be capable to segment the same images into different text-defined regions without retraining, known as open-vocabulary segmentation. For other modalities than RGB images, input-text pairs are not available in the same scale. This work investigates how other modalities can be processed without directly available input-text pairs and how such geometric cues as additional inputs can improve the performance of open-vocabulary segmentation. For this purpose, we adopt a two-stage model, where the first stage generates the mask proposals and proposal embeddings via MaskFormer, while the second encodes masked images and text into the same latent space using CLIP. To adapt this two-stage model for the new modality, we add an additional backbone into MaskFormer and a CLIP encoder into the second stage for the new modality. We conduct experiments with various modality generation approaches, as well as fusion techniques to combine color-based features with geometry-based features in the latent space. We observe an improvement when we only have the additional backbone in the first stage. We conduct an ablation study to adapt CLIP to different modalities. We also show that having the additional geometric cues as input improves the segmentation boundaries and reduces pixel misclassification. We validate our best-model zero-shot on the real-world NYUv2 RGB-D dataset and achieve a 4.5% improvement in mIoU over the original open vocabulary segmentation model.

1. Introduction

Open vocabulary segmentation has emerged for RGB frames to enrich scene understanding through the interpretation of arbitrary queries [4, 6, 8, 9, 15]. It enables rich semantic understanding by not restricting segmentation to predefined categories. This has significant implications for user-facing applications and autonomous robotics.



OV-Seg **MOVSeg (Ours)**

Figure 1: **Multi-Modal Open Vocabulary Segmentation.** We leverage additional input modalities to boost the performance of open-vocabulary image segmentation approaches.

These applications could benefit from the use of geometric cues, which can be complementary to RGB features, as illumination changes do not impact them. Geometric cues are now available in many devices and robots allowing the real-world application of multi-modal open vocabulary segmentation. However, we find a scarcity of large-scale color- and geometry-based input-text pairs. To address this limitation, we employ zero-shot modality generation techniques and adapt models pre-trained on RGB-text pairs to process multi-modal input-text pairs.

In this study, we explore whether off-the-shelf monocular prediction techniques can facilitate learning open vocabulary segmentation. We generate new modalities for a large-scale RGB dataset, COCO-stuff [1] to train our newly-designed model. Next, we directly evaluate our method on ADE20K [17] with synthetic modalities and NYUv2 dataset [13] with noisy depth frames. This shows the cross-dataset transferability of open vocabulary segmentation. We also experiment with different monocular predictors and fusion techniques to fuse RGB-X features, where X is the additional modality. Moreover, we propose adapting CLIP for different modalities to further improve generalizability. We show that the addition of geometric cues improves the segmentation results leading to more accurate predictions and achieving a 4.5% improvement in mIoU values on NYUv2.

2. Related Work

2.1. Modality Generation Methods

Given the lack of available large-scale multi-modal datasets, we use large-scale RGB datasets and generate new modalities using MiDaS [12] and Omnidata Vision [5]. They are shown to work zero-shot on different datasets for modality estimation from RGB frames. In our experiments, we use depth frames generated from MiDaS and Omnidata and the surface normals generated by Omnidata.

2.2. Multi-Modal Fusion

Semantic segmentation from RGB frames can be enhanced by using geometric cues. For closed-set segmentation, it is shown that using geometric cues enhances the segmentation quality [2, 7, 14, 16]. We experiment with simple late-fusion techniques, average, and summation in addition to Attention Complementary Module (ACM) from ACNet [7]. Although the original paper applies ACM throughout the encoding stage, we only use them at the end of encoding to keep the number of learnable parameters smaller and keep the original model structure.

2.3. Open Vocabulary Segmentation

Open vocabulary segmentation is the task of understanding an image and where the objects are located based on arbitrary text queries. After CLIP [11] was proposed, open vocabulary segmentation research has shifted towards aligning pixel- or segment-level embeddings with text embeddings obtained from CLIP. In this direction, LSeg [8] uses CLIP text embeddings and aligns pixel-level features to the corresponding semantic class. OpenSeg [6] instead aligns segment-level features with text embedding via region-word grounding. Our approach is based on the two-stage model approach OVSeg [9], where the first stage generates class-agnostic mask proposals, and the second utilizes CLIP to find the associated class.

Our approach differs from previous methods as we utilize additional geometric cues. We also introduce a separate backbone in MaskFormer for the new modality and perform two-stage training. Moreover, to the best of our knowledge, we are also the first to fine-tune CLIP models to process multi-modal inputs.

3. Methodology

We leverage a two-stage model for open vocabulary segmentation, where the first stage is a MaskFormer [3] and the second is CLIP image and text encoders [11]. The architecture is based on OVSeg [9], which is shown to generalize well across datasets on RGB images. The MaskFormer, which generates proposal embeddings and mask proposals, was previously trained with the following loss.

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + 20 * \mathcal{L}_{mask} + \mathcal{L}_{dice} \tag{1}$$

where $\mathcal{L}_{cls}$ is the cross entropy loss for each class prediction of a masked region, $\mathcal{L}_{mask}$ is the sigmoid focal loss as used in [10], and $\mathcal{L}_{dice}$ is Dice loss to measure the similarity between the predicted region and the ground truth mask based on the overlapping regions as defined in [6].

To process geometric cues, we add an X-backbone to encode the additional modality to the latent space. Initialized from RGB-backbone, we fine-tune X-backbone with the loss defined in Eq. 1 in a supervised manner. We experiment with the following late-fusion techniques to combine RGB and X embeddings in the latent space.

- **Sum:** $\mathbf{f}_{out} = \mathbf{f}_{rgb} + \mathbf{f}_x$
- **Average:** $\mathbf{f}_{out} = (\mathbf{f}_{rgb} + \mathbf{f}_x)/2$
- **Attention:** $\mathbf{f}_{out} = \mathrm{ACM}(\mathbf{f}_x) * \mathbf{f}_{rgb} + \mathrm{ACM}(\mathbf{f}_{rgb}) * \mathbf{f}_x$

where $\mathbf{f}_{rgb}$ and $\mathbf{f}_x$ are the embeddings of RGB and X frames. For more details in inference, please refer to OVSeg [9].

In the second stage, CLIP is adapted for masked modalities while MaskFormer is frozen. To adapt CLIP to process the additional modality, we apply unsupervised training using RGB-X pairs. Initialized from CLIP-RGB, we add a CLIP-X encoder and freeze all other components. We use cosine similarity loss between features obtained from these two separate encoders. As we freeze CLIP-RGB, we anticipate improving the segmentation further by leveraging the full potential of geometric cues while retaining CLIP's generalization capabilities. For inference, we combine the embeddings from CLIP-RGB and -X encoders with a weighted average where weights are given as hyperparameters. The rest of the prediction is the same as in OVSeg.

4. Experiments

4.1. Datasets and Metric

We train our models with COCO-Stuff 171 [1], and validate them on ADE20K-150 and NYUv2 datasets [13, 17].

COCO-Stuff [1]: COCO-Stuff dataset provides 118K RGB training images with pixel-level annotations from indoor and outdoor scenes. The class annotations come from COCO captions, which include 171 distinct classes.

ADE20K-150 [17]: ADE20K dataset includes RGB images from both indoor and outdoor scenes. It has 2000 images for validation. It comprises 150 classes, 110 of which are distinct from COCO-stuff.

NYUv2 [13]: NYUv2 is a real-world RGB-D dataset. It includes a total of 1449 RGB images with 40 classes. We test our approach on NYUv2 to evaluate how well our model performs on a real-world dataset.

To evaluate our models, we use **mean Intersection over Union (mIoU)**, the mean ratio of the intersection and union

Method	Modal-X Fusion	mIoU
OVSeg [9]	-	29.6
OVSeg + D-Backbone	Sum	30.1
OVSeg + D-Backbone	Average	**30.2**
OVSeg + D-Backbone	Attention	29.9

Table 1: **Comparison of RGB-X Fusion Techniques.** We test our model on ADE20K with monocular depth frames generated by MiDaS [12].

of the target ground-truth pixels, and the predictions. It measures the quality of the segmentation masks. Using mIoU metric and evaluation datasets, we validate our methods' segmentation and generalization capabilities.

4.2. MaskFormer Tuning for RGB-X Frames

First, we add a separate X-backbone to the MaskFormer architecture and test different fusion techniques. To do so, we generate the depth frame for each image in COCO-Stuff and ADE20K via MiDaS library. We fine-tune X-backbone with COCO-stuff and validate it over ADE20K.

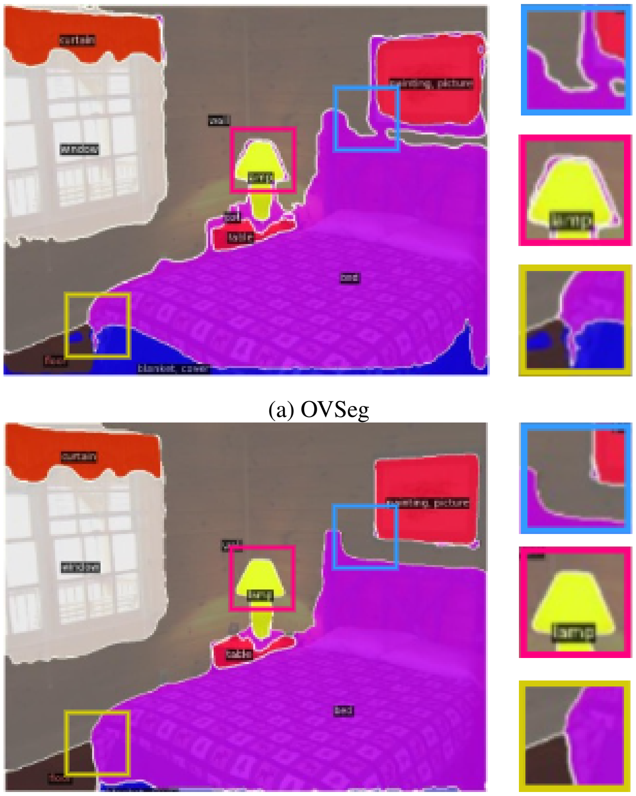
We report the results with different fusion techniques and compare them against OVSeg in Table 1. Adding depth improves the segmentation metrics slightly when tested over ADE20K with synthetic depth frames, leading to more accurate segmentation masks and class labels. When fusion techniques are compared, we see that average achieves the best results. The fusion with attention mechanism does not perform as well as average, which is likely due to additional learnable parameters introduced by the attention model.

We also perform a qualitative comparison when the chosen fusion technique is applied to combine RGB features with the features extracted from Omnidata-generated surface normals. This time, we fine-tune the X-backbone with synthetic surface normals, then test it over ADE20K. As seen in Figure 2, the use of surface normals gives finer segmentation boundaries while also removing some misclassified regions.

4.3. CLIP Adaptation for Geometric Cues

After the first step, we add the CLIP-X encoder for the new modality into the second stage of the model. It is a copy of the original CLIP ViT-L/14 architecture. It encodes the new modality into the same geometric space as RGB features and text features. To retain CLIP's generalization capabilities, we take an unsupervised learning approach, keep the weights of the original CLIP model and only fine-tune the new backbone using similarity loss between RGB-features and X-features.

While training the model, we first evaluate the model with CLIP-X discarding CLIP-RGB embeddings and validating if the new encoder is learning according to the objective. We present our results in Table 2, where we compare different modality generation methods to synthesize the new modality before adapting CLIP.



(a) OVSeg

(b) MOVSeg (Omnidata Surface Normal)

Figure 2: **Qualitative Comparison of Segmentation Masks.** We compare the segmentation masks generated by baseline OVSeg against our multi-modal open-vocabulary segmentation approach MOVSeg.

Patch Embedding	Modal-X Generation	mIoU
CLIP-RGB	-	**29.6**
CLIP-D	MiDaS	9.14
CLIP-D	Omnidata Depth	12.8
CLIP-N	Omnidata Normal	19.4
CLIP-RGB + CLIP-D	MiDaS	29.4
CLIP-RGB + CLIP-D	Omnidata Depth	26.5
CLIP-RGB + CLIP-N	Omnidata Normal	29.3

Table 2: **Ablation on Patch Embeddings.** We compare different patch embeddings in CLIP.

Given the results of Table 2, the CLIP-N has the highest performance when tested alone. This could be due to the information-richness of surface normals compared to depth frames. However, when combined, performance does not improve compared to the baseline model with only RGB images. This might have been caused by the quality of synthetic modalities or because CLIP-RGB was originally trained with 400 million images and it is not trivial to adapt

Model	Modal-X Fusion	mIoU
OVSeg [9]	-	34.2
MOVSeg (Ours)	Average	**38.7**

Table 3: **Results on NYUv2 [13].** We compare OVSeg [9] with our multi-modal open vocabulary segmentation model.

it to new domains with only 118K RGB-X pairs.

4.4. Zero-Shot Segmentation Results on NYUv2

Our objective is to obtain a model that can use geometric cues for improving the open-vocabulary segmentation capabilities. Hence, we test our best model on the real-world multi-modal dataset, NYUv2, and compare it against the original model, which only uses RGB values. We achieve a 4.5% improvement in the mIoU metric (see Table 3).

5. Conclusion

In this paper, we present our findings on open vocabulary segmentation from multi-modal inputs. We experiment with different modality generation methods and fusion techniques. We propose adapting CLIP to new modalities to enhance the segmentation results.

We observe slight improvements in half-synthetic datasets when we process RGB-X frames. Furthermore, we test our model on the real-world dataset NYUv2 achieving a 4.5% improvement in mIoU. This demonstrates the model's ability to enhance performance even when learning from synthetic depth frames.

Future work could involve exploring advanced modal-fusion techniques with shared weights between RGB-X encoders and experimenting with attention modules for CLIP-RGB and CLIP-X fusion. Additionally, the use of triplet loss could be explored for CLIP adaptation.

Limitations

Our study reveals limitations for open vocabulary segmentation on multi-modal frames. We find a scarcity of large-scale real-world datasets with diverse classes, hindering semantic segmentation tasks. Existing modality generation methods are not robust enough for downstream learning, introducing noise and limiting model performance.

Adapting CLIP to new domains poses challenges, given its training on millions of image-text pairs. The modality generation techniques further introduce noise, degrading performance. To address these issues, we could explore simulation-generated high-quality RGB-X pairs for CLIP adaptation and experiment with advanced loss functions, such as triplet loss.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[2] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *ICCV*, 2021.

[3] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

[4] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. 2023.

[5] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, 2021.

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

[7] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *ICIP*, 2019.

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

[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.

[10] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *PAMI*, 2018.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *ICML*, 2021.

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

[13] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[14] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, 2022.

[15] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2023.

[16] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022.

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