Self-guided Semantic Segmentation

University of Amsterdam

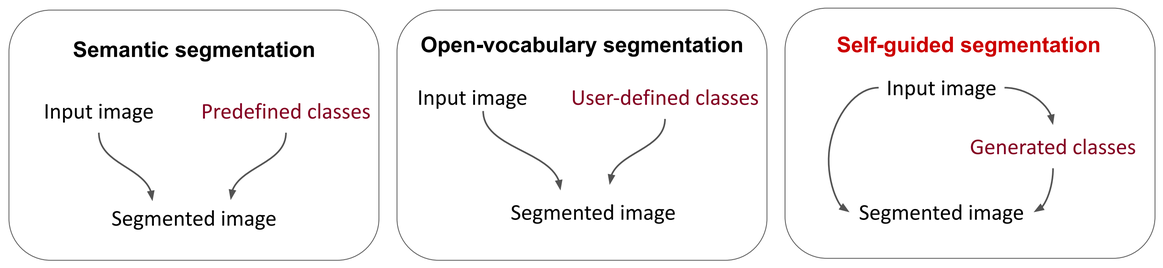
ATLAS LAB.

Maksymilian Kulicki, Osman Ülger, Martin R. Oswald

Novel task*: Self-guided segmentation

- **Semantic segmentation** typically relies on a predefined set of classes.
- **Open-vocabulary segmentation (OVS)** instead relies on classes provided by the user.
- We propose **Self-guided segmentation (SegSeg)**, a novel segmentation task in which the model generates open-vocabulary class names as a part of the process.

*concurrent work with: Rewatbowornwong et al. "Zero-guidance Segmentation Using Zero Segment Labels". ICCV 2023



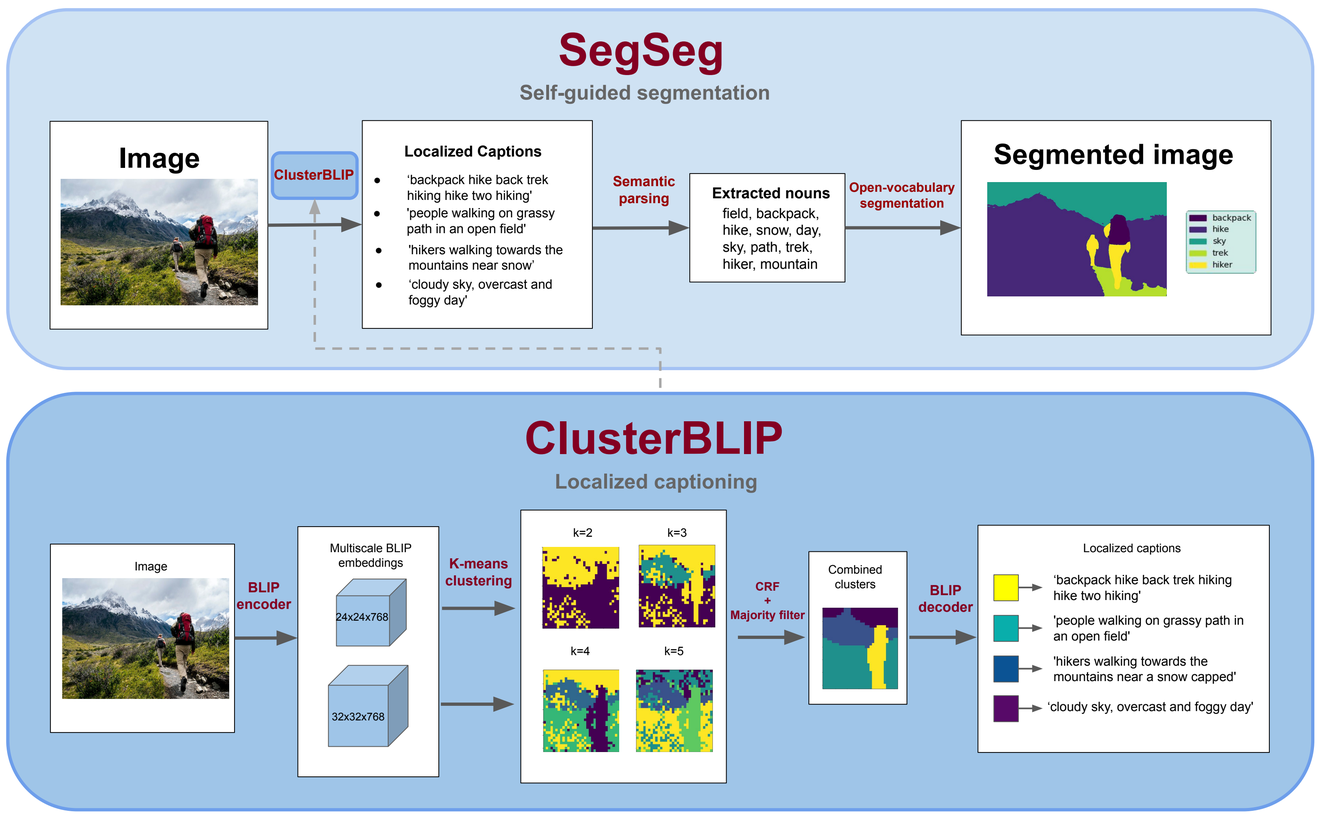
Method

SegSeg combines two pretrained models: BLIP and X-Decoder.

- **BLIP** is a Vision-Language model, containing image encoder and text decoder modules that can be used for image captioning. The captions serve as a source of class names.
- **X-Decoder** is a generalized decoding model capable of various open-vocabulary vision tasks, including OVS.

Image captioning models, including BLIP, tend to focus on the most salient objects and fail to exhaustively describe all elements of a scene. To address that, we introduce **ClusterBLIP**, a method that utilizes pretrained BLIP to generate localized captions.

We use BLIP encoder to generate patch embeddings, which are used to find semantic clusters in the image. The embeddings corresponding to the clusters are used as input for BLIP text decoder to generate localized captions. We parse the captions to extract a list of nouns. These nouns are used as input for X-Decoder, which generates the final segmentation.



Baselines

Since we introduced the task ourselves, there are no established baselines. We compare our method with open-vocabulary segmentation and with more naive self-guided approaches that do not utilize ClusterBLIP.

Open-vocabulary:

- X-Decoder with ground-truth classes present in the image. This setting, with only correct labels provided, serves as upper bound of our method. We use it to calculate $\mathbf{mIoU}_{\mathbf{norm}}$ measuring performance relative to the upper bound
- X-Decoder with all possible ground-truth classes from the dataset

Self-guided:

- BLIP + X-Decoder: caption generation with one BLIP embedding per image
- Grid BLIP + X-Decoder: image divided in a 4-part square grid, one BLIP embedding per part

In addition, we explore generating multiple captions per embedding. This provides a larger and more diverse set of nouns for X-Decoder.

Results

	Self-guided	Nr of captions	mIoU	mIoU_{norm}
X-Decoder (classes from the image)	✗	-	58.6	100.0
X-Decoder (all CityScapes classes)	✗	-	50.2	85.7
SegSeg (ClusterBLIP + X-Decoder)	✓	1	11.0	18.8
		5	23.4	39.9
		15	36.5	62.3
		25	40.1	68.4
		35	39.0	66.6
BLIP + X-Decoder	✓	1	11.1	18.9
		5	17.3	29.5
		15	22.9	39.1
		25	12.6	21.5
		35	17.7	30.2
Grid BLIP + X-Decoder	✓	1	18.4	31.4
		5	22.5	38.4
		15	32.7	55.8
		25	19.3	32.9
		35	32.1	54.8

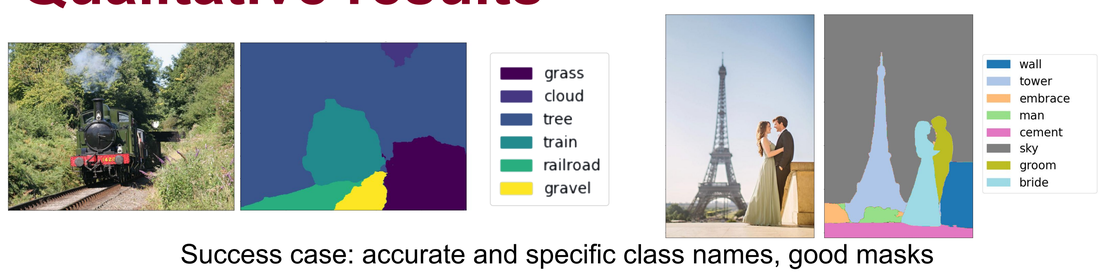
- Our method significantly beats the naive self-guided baselines
- More captions improve performance, the effect saturates around 15-25 captions.
- Our method reaches up to 66.6 percent performance compared to OVS with ground-truth classes provided

Evaluation

We evaluate our method on **CityScapes**. Since our method generates class names in an open manner, we implement a method to map generated labels to classes in the dataset. We use **SentenceBERT** to obtain text embeddings. We measure cosine similarity between embeddings to find the closest match from the dataset for a given label. We reassign the labels and measure the resulting mIoU. The mapping problem is challenging due to its open nature and lack of clear ground truth. Future work is needed to investigate other evaluation approaches.

Example output of our method, with generated label on the left and the closest Cityscapes class on the right. The "taxi/road" mapping is a clear mistake

Qualitative results



Success case: accurate and specific class names, good masks

Failure case 1: Competition of related labels

Failure case 2: good masks, inadequate labels

References

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation", ICML 2022.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. "Generalized decoding for pixel, image and language", CVPR 2023