Open-vocabulary Semantic Segmentation with Frozen Vision-Language Models

Chaofan Ma*[1]
chaofanma@sjtu.edu.cn
Yuhuan Yang*[1]
yangyuhuan@sjtu.edu.cn
Yanfeng Wang†[1,2]
wangyanfeng@sjtu.edu.cn
Ya Zhang[1,2]
ya_zhang@sjtu.edu.cn
Weidi Xie[1,2]
weidi@sjtu.edu.cn

[1] Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
[2] Shanghai AI Laboratory

Abstract

When trained at a sufficient scale, self-supervised learning has exhibited a notable ability to solve a wide range of visual or language understanding tasks. In this paper, we investigate simple, yet effective approaches for adapting the pre-trained foundation models to the downstream task of interest, namely, open-vocabulary semantic segmentation. To this end, we make the following contributions: (i) we introduce **Fusioner**, with a lightweight, transformer-based fusion module, that pairs the *frozen* visual representation with language concept through a handful of image segmentation data. As a consequence, the model gains the capability of zero-shot transfer to segment novel categories; (ii) without loss of generality, we experiment on a broad range of self-supervised models that have been pre-trained with different schemes, *e.g.* visual-only models (MoCo v3, DINO), language-only models (BERT), visual-language model (CLIP), and show that, the proposed fusion approach is effective to any pair of visual and language models, even those pre-trained on a corpus of uni-modal data; (iii) we conduct thorough ablation studies to analyze the critical components in our proposed **Fusioner**, while evaluating on standard benchmarks, *e.g.* PASCAL-5^i and COCO-20^i, it surpasses existing state-of-the-art models by a large margin, despite only being trained on frozen visual and language features; (iv) to measure the model's robustness on learning visual-language correspondence, we further evaluate on a synthetic dataset, named Mosaic-4, where images are constructed by mosaicking the samples from FSS-1000. **Fusioner** demonstrates superior performance over previous models.

1 Introduction

In the recent literature, self-supervised representation learning has made remarkable progress. For example, MoCo [12] and DINO [8] have shown the possibility of learning strong vi-

* Equal contribution. † Corresponding author.

sual representation without using any manual annotation. Noticeably, despite being trained purely by self-supervised learning on images, these image models have shown to implicitly capture the concept of objectness, *i.e.* free semantic segmentation to some extent [34, 44, 48]. Another line of work proposes to learn joint representation for image and text on very large-scale image-text pairs collected from Internet. For example, by training with simple noise contrastive learning, CLIP [38] and ALIGN [24] have demonstrated impressive "zero-shot" transferability and generalizability in various image classification tasks. With the growing computation, it is thus foreseeable that more powerful models will be trained, with even larger scale datasets, further pushing the performance on image classification.

In contrast, semantic segmentation considers a more challenging task, that aims to assign one semantic category to each pixel in an image. By training deep neural networks on large-scale segmentation datasets, recent models have shown great success, for example, FCN [43], U-Net [42], DeepLab [9], DPT [40]. However, the conventional strategy that trains parametric classifiers for individual category, has posed fundamental limitations, as it only allows the model to make predictions on a close-set of predetermined categories at inference time, *i.e.*, the model can only segment objects of the training set classes, and lacks the ability to handle samples from novel (unseen) classes.

In this paper, we explore efficient ways to bridge the powerful pre-trained vision-only, language-only or visual-language models, which, as a result, effectively tackles the problem of **open-vocabulary semantic segmentation**, *i.e.*, segmenting objects from any categories by their textual names or description. In specific, we first encode the input image and category names with *frozen*, self-supervised visual and language models. The computed features are then concatenated and passed to a transformer-based fusion module, that enables the features to be iteratively updated, in condition of the other modality through self-attention. As to obtain the segmentation mask, we further measure the cosine similarity between each pixel and the textual features of each category, *i.e.*, computing dot product between the L2-normalised visual and language embeddings.

To summarize, we make the following contributions: *First*, to tackle the problem of open-vocabulary semantic segmentation, we introduce **Fusioner** with a simple, lightweight transformer-based fusion module, which enables to explicitly bridge powerful, pre-trained visual, language, or visual-language models; *Second*, we demonstrate the idea's effectiveness by experimenting on a wide spectrum of self-supervised models, that are pre-trained with completely different schemes, for example, MoCo v3, DINO, CLIP, BERT, and show that the proposed fusion approach is effective to any pair of visual and language models, even those pre-trained on a corpus of uni-modal data; *Third*, we conduct thorough ablation studies to validate the critical components. Despite all visual and language models are kept frozen, the proposed simple fusion module is able to outperform the existing state-of-the-art approaches in "zero-shot" settings significantly, and is competitive across numerous few-shot benchmarks, *e.g.*, PASCAL, COCO, FSS-1000; *Fourth*, to measure the model's robustness for learning visual-language correspondence, we introduce a new dataset, named Mosaic-4, that can be used in open-vocabulary semantic segmentation to detect whether the models tend to segment saliency that ignore the textual input. **Fusioner** shows superior performance over previous models.

2 Related Work

Pre-trained Vision and Language Models. Self-supervised representation learning has re-

cently made substantial progress in both computer vision and natural language processing. In specific, some pre-training methods adopt contrastive learning (SimCLR [10], MoCo [12], SwAV [7]), metric-learning (BYOL [18], SimSiam [11]) and self-distillation (DINO [8]) can all be seen as powerful feature extractors on downstream tasks. On the other hand, language model pre-training may use a masked language modeling loss (BERT [25], T5 [39]) or next-token prediction loss (GPT [5, 36, 37]). Not surprisingly, large-scale visual-language models have also attracted growing attention [13, 26, 28, 30, 32, 45], a milestone work is Contrastive Language-Image Pre-training (CLIP) [38], that trains on the large-scale image-text pairs with simple noise contrastive learning, and has shown strong capability of aligning two modalities in embedding spaces. Inspired by this work, a series of studies have been proposed to transfer the knowledge of the pre-trained CLIP and extend to various downstream tasks. For example, object detection [17, 19], image captioning [23], referring image segmentation [49], text-driven image manipulation [35], and supervised dense prediction [41], etc. Unlike these works that *fine-tune* CLIP for different downstream tasks, we explore an alternative approach, and show that, simply pairing *any* frozen self-supervised visual and language models with lightweight fusion module, can be a surprisingly strong baseline for open-vocabulary semantic segmentation.

Bridging Pre-trained Models. With the rapid development of foundation models [3, 14, 53], many works have studied effective ways to adapt different downstream tasks by composing different pre-trained models. Frozen [47] proposed to align the pre-trained, frozen language model with vision encoder by learning continuous prompts with only a few examples. Flamingo [1] proposed an architecture that uses large pre-trained vision-only and language-only models to learn a wide range of visual-language task with only limited samples. Socratic Models [54], which uses pre-trained language models, vision-language models, and audio-language models to complete downstream multi-modal tasks, such as image captioning and video-to-text retrieval, without the need of training.

Semantic Segmentation of Novel Categories. To enable a network for segmenting novel categories is still an open and active research question, as most of the existing semantic segmentation methods are limited to a closed set, *i.e.*, the category of test set is the same as the training set. Zero-shot semantic segmentation often take advantage of category-level semantic word embedding to segment novel categories without additional samples. For example, ZS3Net [6], CSRL [29], CaGNet [20], and CaGNet-v2 [21] are generative methods combining a deep visual segmentation model with an approach to synthesize visual features for novel categories based on semantic word embeddings. SPNet [51], JoEm [2], LSeg [27] are discriminative methods mapping each pixel and semantic word to a joint embedding space, and leveraging the joint embedding space to give the class probability. Our approach falls into the latter line, however, we advocate a lightweight fusion module that only aligns the pre-trained, frozen visual and language features.

3 Methods

In this section, we start by formulating the the problem of open-vocabulary semantic segmentation, and then detail our proposed architecture, termed as **Fusioner**, that addresses the task by bridging the pre-trained vision and language models.

Problem Formulation. Following the same setting as defined in LSeg [27], we are given a training set $\mathcal{D}_{\text{train}} = \{(\mathcal{X}, \mathcal{Y}, \mathcal{C}) | \mathcal{X} \in \mathbb{R}^{H \times W \times 3}, \mathcal{Y} \in \mathbb{R}^{H \times W \times |\mathcal{C}|}, \mathcal{C} \subseteq \mathcal{S}\}$, where $\mathcal{X}$ denotes

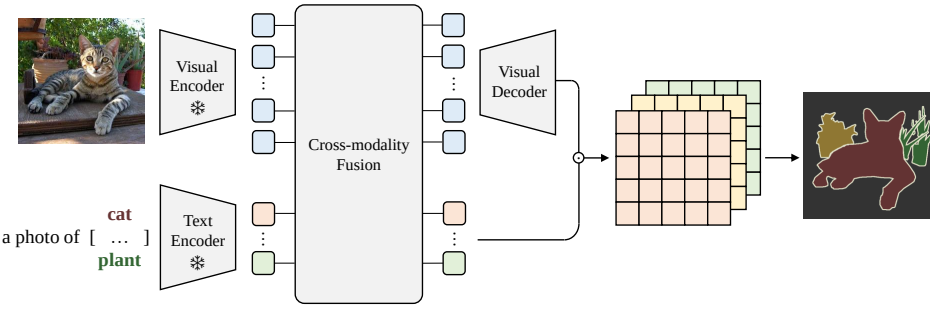


Figure 1. Overview of our proposed **Fusioner**, which consists of a frozen visual and text encoder, a cross-modality fusion module and a visual decoder. The frozen encoders extract features for different modalities, and the fusion module bridges these embedding spaces. After upsampling the visual feature to its original resolution by the visual decoder, segmentation can be acquired by simple computing the similarity between the visual and language modalities.

any input image; $\mathcal{Y}$ refers to the segmentation masks for $\mathcal{X}$ with one-hot encoding; $\mathcal{C}$ refers to a set of seen categories in $\mathcal{X}$; $\mathcal{S}$ is all training (seen) categories in this training set. Our goal here is to train a segmentation model that can partition a test image into semantically meaningful regions of *unseen* categories:

$$\mathcal{Y}_j = \Phi_{\text{FUSIONER}}(\mathcal{X}_j, \mathcal{W}_j) = \Phi_{\text{DEC}}(\Phi_{\text{FUSE}}(\Phi_{\text{VISUAL-ENC}}(\mathcal{X}_j), \Phi_{\text{TEXT-ENC}}(\mathcal{W}_j))), \quad (1)$$

where $\mathcal{W}_j = \{w_j^1, \ldots, w_j^{|\mathcal{W}_j|}\}$ is the categories of interests in one image $\mathcal{X}_j$ ($|\mathcal{W}_j|$ textual words, *e.g.*, "cat", "plant"), and is dynamic for different images. The corresponding $|\mathcal{W}_j|$ output binary segmentation maps denote as $\mathcal{Y}_j \in \mathbb{R}^{H \times W \times |\mathcal{W}_j|}$. Note that, for the conventional close-set segmentation, $\mathcal{W}_j \subseteq \mathcal{S}$, while for the open-vocabulary problem, $\mathcal{W}_j \bigcap \mathcal{S} = \emptyset$, *i.e.*, we evaluate the performance on the novel (unseen) categories that do not appear in the training categories $\mathcal{S}$.

3.1 Architecture

The overall framework of our proposed **Fusioner** is illustrated in Figure 1. It consists of three components: visual and language encoders for extracting features (Section 3.1.1); a cross-modality fusion module to bridge the embedding spaces (Section 3.1.2); and an image decoder that upsamples the visual features to facilitate segmentation on original resolution as input images, and segmentation can be acquired by simple computing the cosine similarity between the visual and language (Section 3.1.3).

3.1.1 Visual and Language Representation

Here, we adopt pre-trained vision and language models as our encoders, and keep them frozen during training. In specific, we mostly consider the transformer-based architectures, due to their good performance, and flexibility for encoding signals of different modalities.

Visual Feature Embeddings. As the input to a vision transformer [15], the image $\mathcal{X} \in \mathbb{R}^{H \times W \times 3}$ is first split into a set of 2D non-overlapping patches, and projected into a sequence of vector embeddings. After adding positional embeddings (inherited from the pre-trained

vision transformer), these tokens are further processed by a series of transformer encoder layers, each layer consists of multi-head self-attention (MHSA) and feed-forward network (FFN) together with layer normalization and residual connections:

$$\mathcal{F}_{\mathrm{V}} = \Phi_{\text{VISUAL-ENC}}(\mathcal{X}) \in \mathbb{R}^{(hw)\times d_i}, \tag{2}$$

where $h = H/p$ and $w = W/p$, p is the patch size, and d_i is the visual embedding dimension. In later experiments, we adopt various popular transformer-based image encoders, that were pre-trained with different self-supervised learning regime, for example, MoCo v3 [12], DINO [8], and CLIP [38].

Textual Feature Embeddings. The text embeddings are generated by encoding the semantic categories $\mathcal{W}$ through a text encoder:

$$\mathcal{F}_{\mathrm{W}} = \Phi_{\text{TEXT-ENC}}(\mathcal{W}) \in \mathbb{R}^{|\mathcal{W}|\times d_w}, \tag{3}$$

where $|\mathcal{W}|$ refers to the input number of categories, and d_w is the dimension of word embeddings. Prior to feeding the semantic category into text encoder, we use multiple prompt templates as decorations, *e.g.*, "a photo of {category} in the scene", and average the output embeddings from text encoder. The complete prompt templates are listed in supplementary. We consider various self-supervised language models that were trained on a large corpus of documents or images as the text encoder, for example, BERT [25], or CLIP [38].

3.1.2 Cross-modality Fusion

Given the visual features $\mathcal{F}_{\mathrm{V}}$ and textual features $\mathcal{F}_{\mathrm{W}}$, we firstly unify the channel dimensions for both visual and textual embeddings by using MLPs, *i.e.* $\mathcal{F}_{\mathrm{V}} \in \mathbb{R}^{(hw)\times d}$, $\mathcal{F}_{\mathrm{W}} \in \mathbb{R}^{|\mathcal{W}|\times d}$, and pass the resulting features through a cross-modality fusion module to adaptively capture the interactions between visual and language signals:

$$[\mathcal{F}'_{\mathrm{V}},\ \mathcal{F}'_{\mathrm{W}}] = \Phi_{\text{FUSE}}([\mathcal{F}_{\mathrm{V}},\ \mathcal{F}_{\mathrm{W}}]), \tag{4}$$

where $[\cdot,\cdot]$ indicates feature concatenation of the visual and textual sequence. Φ_{FUSE} is consisted of multiple transformer encoder layers, effectively capturing the long-range dependencies between the images and associated texts. The multi-modality visual feature $\mathcal{F}'_{\mathrm{V}}$ and textual feature $\mathcal{F}'_{\mathrm{W}}$ have the same shape as $\mathcal{F}_{\mathrm{V}}$ and $\mathcal{F}_{\mathrm{W}}$, and both features are enriched and refined by iteratively attending the other modality.

3.1.3 Visual Decoder

Modality-maintained Upsampling. Here, the visual features are progressively upsampled to the same resolution as the original image, in detail, we first reshape the sequence of visual vectors into a spatial feature map, and upsample it by alternating convolutional and upsampling layers, obtaining high resolution feature maps, *i.e.*, $\mathcal{F}'_{\mathrm{V}} \in \mathbb{R}^{H\times W\times d}$.

Calculating Segmentation Mask. The logits $\hat{y}$ is generated by computing the cosine similarity between the upsampled feature map and textual feature, *i.e.*, $\hat{y} = \mathcal{F}'_{\mathrm{V}} \cdot \mathcal{F}'^{\mathrm{T}}_{\mathrm{W}} \in \mathbb{R}^{H\times W\times|\mathcal{W}|}$, where $\mathcal{F}'_{\mathrm{V}} \in \mathbb{R}^{H\times W\times d}$, $\mathcal{F}'_{\mathrm{W}} \in \mathbb{R}^{|\mathcal{W}|\times d}$, denoting the L2-normalised visual and textual features, and $|\mathcal{W}|$ is the number of categories (textual words). Seen as binary segmentation for each category, the final predictions can be obtained by simply applying sigmoid with a temperature τ and threshold classwise.

3.1.4 Discussion

The closest work to ours is LSeg [27], which also considers to tackle the problem of open-vocabulary semantic segmentation, by explicitly pairing high-capacity image and text encoder, however, there remains **three** critical differences, in LSeg, (i) the visual model (dense prediction transformers [40]) is pre-trained with supervised learning, while we use self-supervised models that can be easily scaled up; (ii) the visual model is optimised end-to-end for segmentation on certain categories, which may potentially lead to the catastrophic forgetting; (iii) the visual and language representation are computed independently with dual encoders, and only fused at the last layer for computing semantic segmentation, thus referring to as a **late fusion**. Such late fusion can potentially suffer from lexical ambiguities, for example, same word (synonym) may refer completely different visual patterns, while **early fusion** (ours) allows to update the features in condition to the other, potentially enabling to learn better visual-language correspondence. In Section 4, we have conducted experiments to validate the superiority of early fusion.

4 Experiment

4.1 Experimental Setups

Datasets. In accordance to prior work on open-vocabulary semantic segmentation [27], we also evaluate our model on two benchmarks: PASCAL-5^i and COCO-20^i. PASCAL-5^i is the extension of PASCAL VOC 2012 [16] with extra annotations from SDS [23], consisting of 20 semantic categories that are divided evenly into 4 folds containing 5 classes each, *i.e.*, $\{5^i\}_{i=0}^3$. COCO-20^i is built from MS COCO [31] and contains 80 semantic categories that are also divided into 4 folds, *i.e.*, $\{20^i\}_{i=0}^3$, with each fold having 20 categories. For each dataset, we conduct 4-fold cross-validation with same hyperparameter setup.

In addition, for robustness test, we introduce a new dataset with images constructed by mosaicking images from FSS-1000 [50], termed as Mosaic-4. FSS-1000 contains pixel-wise annotation of 1000 classes with 10 object-centric images each, in which 240 classes (2400 images) are reserved for test. Mosaic-4 reorganizes the test split of FSS-1000 by randomly sampling 4 images of different categories without replacement, and mosaicking them into one, creating a test list of 600 compound images with explicit distractors. An example can be seen in Figure 2 (a) and (b), where each color represents an individual category.

Implementation Details. We experiment with three different pre-trained vision/language models: the vision-only model (MoCo v3, DINO), the language-only model (BERT), and the vision-language model (CLIP). **Note that**, all of them were kept frozen during training. The cross-modality fusion module contains 12 layers with 8 heads, and the visual decoder consists of k layers of convolution followed by a $2\times$ upsampling, where $k = 4$ for ViT backbone and $k = 5$ for ResNet. We adopt AdamW optimizer, and the learning rate is ramped up during the first 10 epochs to 0.001 linearly. After this warmup, we decay the learning rate with a cosine schedule. The temperature factor $\tau = 0.07$, and cross entropy is used for training.

Evaluation Metrics. We adopt class mean intersection-over-union (mIoU) as our main evaluation metric, The mIoU averages IoU over all classes in a fold: $\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\text{IoU}_c$, where C is the number of classes in the target fold, and IoU_c is the intersection over union of class c. We also consider FB-IoU in some experiments, however, FB-IoU only cares about

the performance on target and non-target regions instead of differentiating categories, where only the foreground and background are considered as two categories ($C = 2$).

4.2 Ability to Bridge Different Pre-trained Backbones

Model	Visual Encoder	Text Encoder	Early Fusion					Late Fusion				
			5^0	5^1	5^2	5^3	mIoU	5^0	5^1	5^2	5^3	mIoU
A	CLIP-B	CLIP-B	50.2	62.4	**51.5**	44.4	52.1	46.3	55.2	36.4	36.7	43.6
		BERT-B	46.6	61.3	44.6	43.7	49.1	47.5	58.5	38.5	39.1	45.9
		BERT-L	48.5	59.3	47.5	43.8	49.7	47.0	56.6	41.4	38.8	46.0
B	CLIP-L	CLIP-L	**61.9**	**70.0**	51.2	52.7	**59.0**	52.6	58.4	45.9	43.7	50.1
		BERT-B	56.9	66.0	45.9	49.6	54.6	56.5	60.2	44.6	47.0	52.1
		BERT-L	56.0	64.4	43.7	52.0	54.0	55.3	60.5	46.3	46.3	52.1
C	DINO-B	CLIP-B	56.4	67.1	49.8	47.5	55.2	57.1	64.6	48.5	46.2	54.1
		BERT-B	56.7	65.1	48.5	45.2	53.9	54.2	61.4	48.7	48.8	53.3
		BERT-L	56.1	65.8	48.7	44.2	53.7	57.4	65.1	48.5	46.3	54.3
D	MoCo-B	CLIP-B	58.9	67.0	47.7	51.7	56.3	56.4	64.9	47.1	49.5	54.5
		BERT-B	59.7	65.3	47.3	**53.4**	56.4	57.4	66.1	46.6	51.2	55.3
		BERT-L	59.1	65.7	49.5	53.1	56.8	54.9	65.0	49.8	50.6	55.0

Table 1. Ability of bridging different pre-trained backbones on PASCAL-5^i. CLIP-B (or -L) means CLIP image/text encoders using ViT-B (or -L); BERT-B (or -L) means BERT-Base (or -Large); DINO-B or MoCo-B means using ViT-B backbone, respectively.

As illustrated in Table 1, **early fusion** refers to our proposed fusion approach, while **late fusion** denotes similar the idea as LSeg [27], where the visual or language features are separately processed with 6 MLPs layers, and only fused at the last segmentation layer.

Here, we can make three observations: (i) pairing the frozen visual and language models can be surprisingly powerful, even for models that are pre-trained on a corpus of uni-modal data, for example, in model-D, with MoCo-B as visual encoder, and BERT-L as language encoder; (ii) early fusion consistently outperforms the late fusion, that validates the conjecture that visual-language correspondence can be better captured by allowing the feature of one modality to be updated in condition to the other; (iii) the segmentation performance tends to improve with the model scale, for example, the model-B-CLIP works significantly better than model-A-CLIP, despite both are pre-trained with image-text pairs. For latter experiments, we adopt the pair in model-B with both encoders CLIP-L for the best performance.

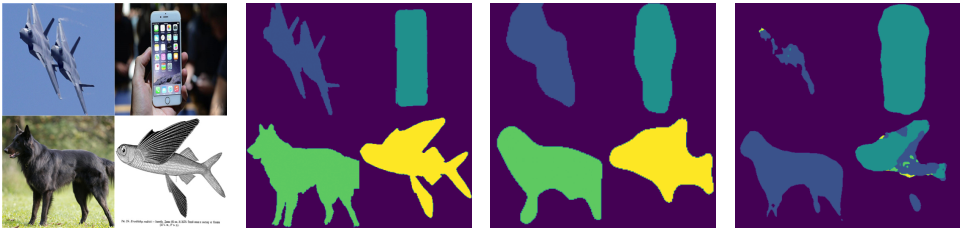
4.3 Comparison with State-of-the-art

Following LSeg [27], we also compare our method with various open vocabulary segmentation methods: ZS3Net [6], SPNet [51] and LSeg [27] on PASCAL-5^i and COCO-20^i.

As shown in Table 2, our proposed **Fusioner** with pre-trained frozen visual-language models (CLIP-L) achieves state-of-the-art results on both PASCAL-5^i and COCO-20^i, outperforming LSeg [27] by a significant margin on mIoU. In contrast to PASCAL-5^i, COCO-20^i is larger in scale and richer in objects, for example, there may exist over 5 categories in one image, making it much more challenging. For late fusion models such as LSeg, the visual feature will be dominated by the most salient objects, however, our cross-modality fusion module can interchange information between visual and language features, adapt each other iteratively. This may explain the performance gap between us and LSeg.

Model	Backbone	PASCAL-5^i					COCO-20^i				
		5^0	5^1	5^2	5^3	mIoU	20^0	20^1	20^2	20^3	mIoU
SPNet [51]	ResNet101	23.8	17.0	14.1	18.3	18.3	-	-	-	-	-
ZS3Net [3]	ResNet101	40.8	39.4	39.3	33.6	38.3	18.8	20.1	24.8	20.5	21.1
LSeg [27]	ResNet101	52.8	53.8	44.4	38.5	47.4	22.1	25.1	24.9	21.5	23.4
LSeg [27]	ViT-L/16	61.3	63.6	43.1	41.0	52.3	28.1	27.5	30.0	23.2	27.2
Fusioner (ours)	ResNet101	46.8	56.0	42.2	40.7	46.4	26.7	34.1	26.3	23.4	27.6
Fusioner (ours)	ViT-L/14	**61.9**	**70.0**	**51.2**	**52.7**	**59.0**	**31.7**	**35.7**	**34.9**	**31.8**	**33.5**

Table 2. Comparison of mIoU on PASCAL-5^i and COCO-20^i.



(a) Example image (b) Ground-truth (c) Our prediction (d) LSeg prediction

Figure 2. An example image and ground-truth of Mosaic-4 dataset, with models' predictions. The same color indicates the same category, *i.e.* "stealth_aircraft", "iphone", "groenendael" and "abe's_flyingfish" from top-left to bottom-right. Our model can distinguish different categories inputs compared with LSeg. Best viewed in color.

4.4 Robustness on Mosaic-4

Here, we measure the robustness of **Fusioner** trained on FSS-1000 using our synthesized Mosaic-4 dataset. We take $|\mathcal{W}| = 4$, *i.e.*, 4 category embeddings as input for one image, during training and testing, and generate 4 binary prediction for each category. For LSeg, we input 4 categories together with text "others" representing the background, and conduct a pixel-wise classification into 5 categories. As shown in Table 3, our model can significantly outperform LSeg. In Figure 2, our model can better differentiate different textual inputs along with their corresponding mask. However, LSeg is confused about "groenendael" and "abe's_flyingfish", and segmenting "stealth_aircraft" and "iphone" with large false positives.

Model	mIoU	FB-IoU
LSeg [27]	19.5	58.2
Fusioner (ours)	**53.7**	**76.3**

Table 3. Results on Mosaic-4.

Model	. Fusion	Decoder	5^0	5^1	5^2	5^3	mIoU
B0-CLIP-L	✓	✓	**61.9**	**70.0**	**51.2**	**52.7**	**59.0**
B1-CLIP-L	✓	✗	59.4	63.7	47.3	44.0	53.6
B2-CLIP-L	✗	✓	48.3	54.4	41.4	42.2	46.6
B3-CLIP-L	✗	✗	16.7	21.7	20.0	20.9	19.8

Table 4. Ablation study on PASCAL-5^i.

4.5 Ablation Study

To investigate the importance of each component in **Fusioner**, we conduct ablation studies on the cross-modality fusion and the visual decoder, and change one variable at a time. All experiments are based on the best model in Table 1, namely, model-B-CLIP-L. As illustrated in Table 4, model-B0-CLIP-L with both fusion and decoder gives the best results. Directly

upsampling the visual feature without visual decoder leads to about a 5% decline, while breaking the connection between visual and text modalities by skipping the fusion module results a sharp drop of 21% in performance, which demonstrates the necessity of our cross-modality fusion module. However, when neither the fusion module nor the visual decoder is applied, no trainable parameters are introduce in the entire pipeline, which, unsurprisingly, yields the poorest result.

4.6 Transferability to Other Datasets

Ideally, open-vocabulary semantic segmentation should be able to handle any textual label regardless of the domain shift between different datasets. Here we evaluate on a more generalizable setting, that is, to test our COCO-trained model on PASCAL VOC following the work of [4]. As shown in Table 5, 20^i means the model was trained on fold i of COCO-20^i and tested on the whole PASCAL VOC dataset, after removing the seen classes in corresponding training split. Details can be found in supplementary. For evaluation, in addition to LSeg [27], we also compare the latest few-shot method RPMM [52], CWT [53], and PFENet [46]. It can be seen from Table 5 that our method outperforms the previous state-of-the-art open-vocabulary method and is comparable to various few-shot methods.

Model	Backbone	Method	20^0	20^1	20^2	20^3	mIoU
RPMM [52]	ResNet50	5-shot	40.2	58.0	55.2	61.8	53.8
PFENet [46]		5-shot	45.1	**66.8**	**68.5**	**73.1**	63.4
CWT [53]		5-shot	**60.3**	65.8	67.1	72.8	**66.5**
RPMM [52]	ResNet50	1-shot	36.3	55.0	52.5	54.6	49.6
PFENet [46]		1-shot	43.2	**65.1**	**66.5**	**69.7**	**61.1**
CWT [53]		1-shot	**53.5**	59.2	60.2	64.9	59.5
LSeg [27]	ResNet101	zero-shot	24.6	-	34.7	35.9	31.7
Fusioner (ours)	ResNet101	zero-shot	31.0	53.7	41.7	51.3	44.4
Fusioner (ours)	ViT-L	zero-shot	**39.9**	**70.7**	**47.8**	**67.6**	**56.5**

Table 5. Transferability from COCO-20^i to PASCAL VOC. Here, LSeg results are generated by averaging the three officially released checkpoints (no ViT-L backbone, only fold 0,2,3 for ResNet).

4.7 Qualitative Results

In Figure 3, we show the qualitative results for our model on open-vocabulary segmentation. Specifically, the subimages in each row include the original image, and segmentation results of seen and unseen (marked as *) categories. As can be seen, our proposed **Fusioner** can successfully predict the unseen categories.

5 Conclusion

With the growing interest in Foundation Models [3], we believe it will be of great significance for the community, to efficiently adapt these powerful vision and language models for the downstream task of interest. Here, we introduce **Fusioner**, a simple, lightweight cross-modality fusion module, that explicitly bridged a variety of self-supervised pre-trained visual/language models for open-vocabulary semantic segmentation. We evaluate on two standard benchmarks (PASCAL and COCO), and conduct thorough ablation studies to demonstrate the effectiveness of our model. Despite the simplicity of the proposed idea, we demonstrate state-of-the-art on all standard benchmarks.

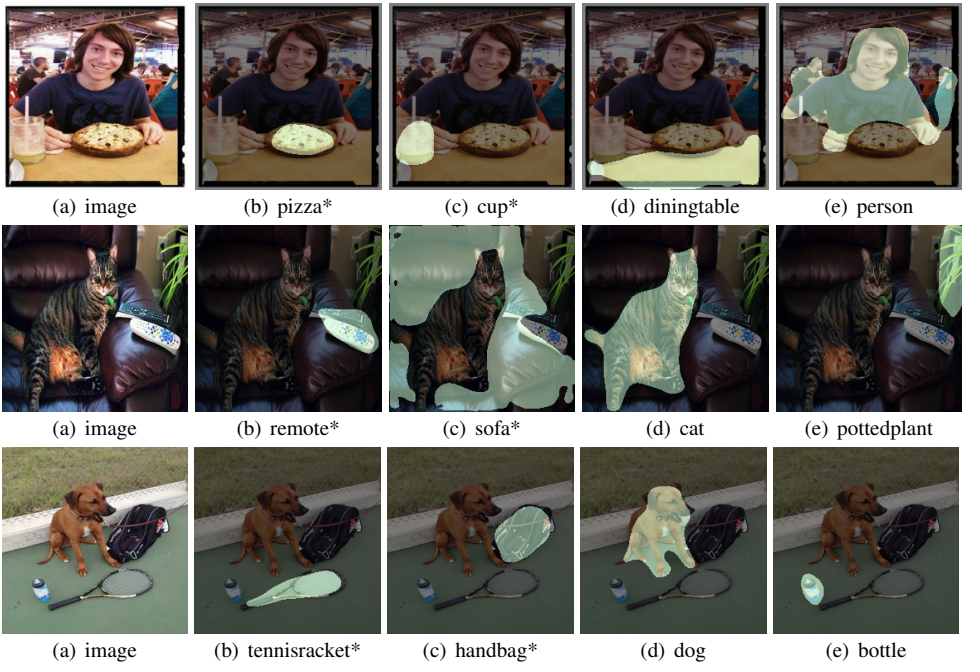


Figure 3. Qualitative results on COCO-20^i. (a) input images, (b)-(e) segmentation masks for different categories. Unseen categories are marked as *.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2020YFB1406801), National Natural Science Foundation of China (62271308), 111 plan (No. BP0719010), and STCSM (No. 18DZ2270700, No. 22511105700), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[2] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9536–9545, 2021.

[3] Rishi Bommasani and et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[6] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[14] Aakanksha Chowdhery and et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[17] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Expand your detector vocabulary with uncurated images. In *European Conference on Computer Vision*, 2022.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[19] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

[20] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020.

[21] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[22] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 991–998. IEEE, 2011.

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[24] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[25] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[27] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

[28] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11336–11344, 2020.

[29] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 33:10317–10327, 2020.

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8741–8750, 2021.

[34] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[38] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

[40] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12159–12168, 2021.

[41] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[43] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:640–651, 2017.

[44] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *IEEE Conference on Computer Vision and Pattern Recognition L3DIVU Workshop*, 2022.

[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[46] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[47] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[48] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[49] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022.

[50] Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2866–2875, 2020.

[51] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019.

[52] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 763–778. Springer, 2020.

[53] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[54] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.