MTA-CLIP: Language-Guided Semantic Segmentation with Mask-Text Alignment

Anurag Das, Xinting Hu, Li Jiang, and Bernt Schiele

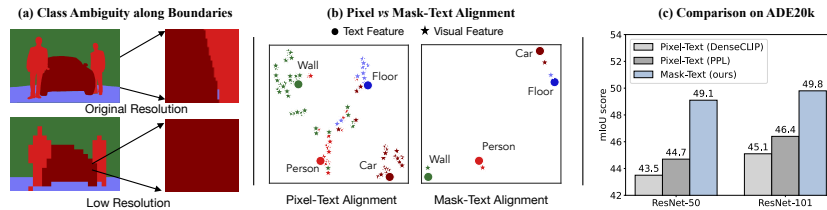Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

Fig. 1: (a): Pixel-Text Alignment relies on low-resolution image features for alignment and often encounters class boundary ambiguities. (b): t-sne visualisation for Pixel-Text *vs.* Mask-Text Alignment. Pixel features of "Wall" are misaligned with the text feature of "Person" for Pixel-Text Alignment. (c): Comparative results on the ADE20k, showcase the superior performance of Mask-Text Alignment in semantic segmentation tasks. on the ADE20k, showcase the superior performance of Mask-Text Alignment over Pixel-Text Alignment methods in semantic segmentation tasks.

Abstract. Recent approaches have shown that large-scale vision-language models such as CLIP can improve semantic segmentation performance. These methods typically aim for pixel-level vision-language alignment, but often rely on low-resolution image features from CLIP, resulting in class ambiguities along boundaries. Moreover, the global scene representations in CLIP text embeddings do not directly correlate with the local and detailed pixel-level features, making meaningful alignment more difficult. To address these limitations, we introduce MTA-CLIP, a novel framework employing mask-level vision-language alignment. Specifically, we first propose Mask-Text Decoder that enhances the mask representations using rich textual data with the CLIP language model. Subsequently, it aligns mask representations with text embeddings using Mask-to-Text Contrastive Learning. Furthermore, we introduce Mask-Text Prompt Learning, utilizing multiple context-specific prompts for text embeddings to capture diverse class representations across masks. Overall, MTA-CLIP achieves state-of-the-art, surpassing prior works by an average of 2.8% and 1.3% on standard benchmark datasets, ADE20k and Cityscapes, respectively.

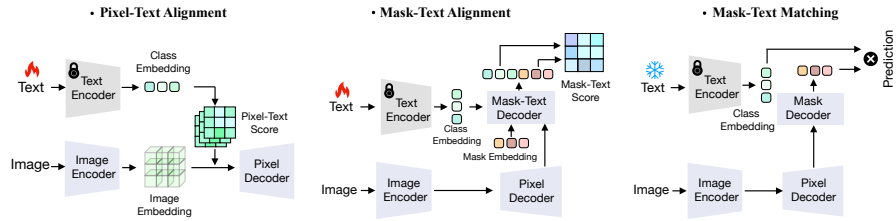**Keywords:** Scene Understanding · Vision Language Models

Fig. 2: Comparison between Pixel-Text Alignment, Mask-Text Alignment (ours) and Mask-Text Matching segmentation architectures. Flames and snowflakes refer to learnable and frozen parameters, respectively.

1 Introduction

Semantic segmentation, a fundamental task in computer vision, has been significantly advanced by the integration of language models, particularly with the advent of large-scale vision-language models (VLMs) such as CLIP [30]. Recent studies [20,31] have demonstrated that pre-training with CLIP can considerably enhance the performance of semantic segmentation. The core strength of CLIP is its training on diverse datasets rich in visual and textual content, which enables a context-sensitive understanding of visual content. In semantic segmentation, the adoption of CLIP in training is defining new benchmarks, highlighting the benefits of cross-modal embeddings.

The common practice of integrating CLIP into semantic segmentation adopts pixel-text alignment [20,31]. In this approach, each pixel feature is aligned with the text embedding of its corresponding ground truth class label as illustrated in Fig. 2 (a). However, this approach faces several challenges. Firstly, due to computational constraints, pixel-text alignment often operates on low-resolution image features. This down-sampling of image features, caused by the network's image encoder stride and pooling, often results in the blurring of boundaries between different classes. This results in inaccurate alignment with the respective text embeddings. As illustrated in Fig. 1 (a), a region containing a "car" and a "person" is indiscriminately labeled as "car" after down-sampling. As a result, the text embedding for "car" does not accurately align with the pixel-level image feature over that region. Moreover, CLIP performs image-text matching [13], resulting in text embeddings with global image-level semantics. It introduces mismatches when the text embeddings are aligned with the detailed and localized pixel features. As illustrated in Fig. 1 (b), certain pixel features for "wall" can be easily confused with the text embedding of "person". In contrast, when using mask features[1] for classes "wall" and "person", the mask features align more effectively with the corresponding text embeddings.

To address the challenges posed by pixel-text alignment, our MTA-CLIP introduces the Mask-Text alignment. Unlike pixel representations, each mask

---

[1] we observe only one mask feature captures the entire class, for each of the classes in this sample. We provide more visualisations in the supplement.

representation with inherent boundary information captures entire entities or regions within a class. Such embeddings are more in line with CLIP's pre-training approach and avoid class ambiguity compared to per-pixel alignment. The key component of MTA-CLIP is the novel Mask-Text Decoder. Different from Mask-Text matching (Fig. 2 [11, 14, 40, 41, 49]) that keeps the text embeddings fixed, our Mask-Text Decoder updates both mask and text embeddings. This enhances masks with rich textual data while enriching text embeddings with visual context using layer-wise attention mechanism. Furthermore, MTA-CLIP uses a contrastive learning approach within the decoder. This approach aligns the enhanced mask embeddings with corresponding text embeddings back into the CLIP feature space, leveraging the potential of CLIP's pre-trained model to regularize the mask representations. Additionally, MTA-CLIP introduces Mask-Text Prompt Learning to effectively manage the diverse contexts within mask segments of the same class. This method utilizes distinct prompts for varying scenarios, thereby enhancing the model's adaptability and contextual awareness.

Overall, our MTA-CLIP achieves a Mask-Text alignment that addresses the inherent limitations of pixel-text alignment. Our contributions are summarized as follows:

- We present MTA-CLIP, a novel CLIP based semantic segmentation framework that shifts from pixel-text to mask-text vision-language alignment, aligning more closely with CLIP's holistic scene understanding and effectively resolving class ambiguities at boundaries.
- MTA-CLIP introduces Mask-Text Alignment comprising of two key components, Mask-Text Decoder, which enhances and aligns mask representations with rich textual data, and Mask-Text Prompt Learning, which employs multiple context-specific prompts to capture diverse class contexts within masks.
- Extensive evaluation on standard datasets like ADE20k and Cityscapes shows that MTA-CLIP outperforms existing state-of-the-art methods, establishing new benchmarks in the semantic segmentation task with Mask-Text vision-language alignment.

2   Related Work

Semantic Segmentation. Semantic segmentation involves the task of assigning a categorical label to each pixel within an image, a challenge that has been extensively addressed using Deep Neural Networks (DNNs) [15, 19, 24, 32]. Previous research has predominantly concentrated on two architectural paradigms, namely Convolutional Neural Networks (CNNs) and Transformer-based models. CNNs [2, 3, 26, 29, 42, 45], with their ability to capture local spatial hierarchies have shown to be highly effective. Noteworthy methods within this domain emphasize the utilization of multi-scale features, exemplified by DeepLab's Atrous Spatial Pyramid Pooling (ASPP) [2, 15], PSPNet's Pyramid Pooling Module (PPM) [45], Semantic Feature Pyramid Network (FPN) [18] with its

Feature Pyramid Network, incorporation of shape information as an additional prior [1,10,34,44], and integration of high-resolution features [36,43]. Conversely, transformers [5,33,35,39], distinguished by their capacity to capture global context, have exhibited superiority over CNN counterparts for semantic segmentation. While the majority of prior works have focused on per-pixel segmentation, assigning a label to each individual pixel, our work aligns with mask-based segmentation networks [5,6], aiming to partition an image into distinct masks with associated labels. Differently, we further enhance the mask representations with rich textual data following Mask-Text Alignment.

Open Vocabulary Semantic Segmentation. To segment potentially unlimited classes, open vocabulary segmentation methods "align" pixel [7,22] or mask features [11,14,40,41,49] with textual representations. We refer open-vocabulary based "alignment" as "matching", given that the text features are not altered in such methods. Pixel-Text Matching method [7], introduced a cost-aggregation that improves CLIP finetuning generalisation for open vocabulary segmentation. Mask-Text Matching (Fig. 2) methods [11,14,41,47], propose a two-stage framework, where mask proposals are generated in the first stage, and are matched with CLIP text features in the next stage. Differently, we focus on closed-set semantic segmentation, where our "Mask-Text Alignment" enhances both text and visual features, essential for optimal performance (see supplement Tab. 5). Further, we also learn multiple contextual prompts for text embeddings to best align diverse class representations across masks, whereas the prior works [11,14,41,47] do not use learnable prompts and use the handcrafted prompts such as "a photo of", "a painting of" instead.

Language Guided Semantic Segmentation. Recent works [20, 31] have demonstrated that leveraging large-scale pretrained Vision-Language Models (VLMs) such as CLIP [30] surpasses ImageNet-based approaches, attributed to their enhanced transferability to downstream tasks. DenseCLIP [31] initially proposed an extension of CLIP for downstream semantic segmentation tasks employing a pixel-text alignment method. Subsequently, PPL [20] expanded upon DenseCLIP by introducing additional probabilistic prompts designed to capture diverse attributes in the data. Notably, both DenseCLIP [31] and PPL [20] rely on pixel-text vision language alignment, particularly at lower resolution features. Diverging from their approach, our proposal introduces a Mask-Text vision language alignment, specifically addressing challenges associated with pixel-text alignment (see Sec. 1).

Prompt Learning. Adapting Vision-Language Models (VLMs) to downstream tasks involves the finetuning of the model, a process that can result in the degradation of learned representations and potentially impact model transferability. Moreover, with an increase in model size, finetuning becomes computationally expensive. The concept of prompt learning, as introduced in prior studies such as [16,17,21,25], proposes an alternative approach by acquiring additional prompt tokens that tailor the large-scale VLM to specific downstream tasks. CoOp [51] optimizes the Contrastive Language-Image Pretraining (CLIP) model for downstream tasks by acquiring a set of input prompts for the VLM's
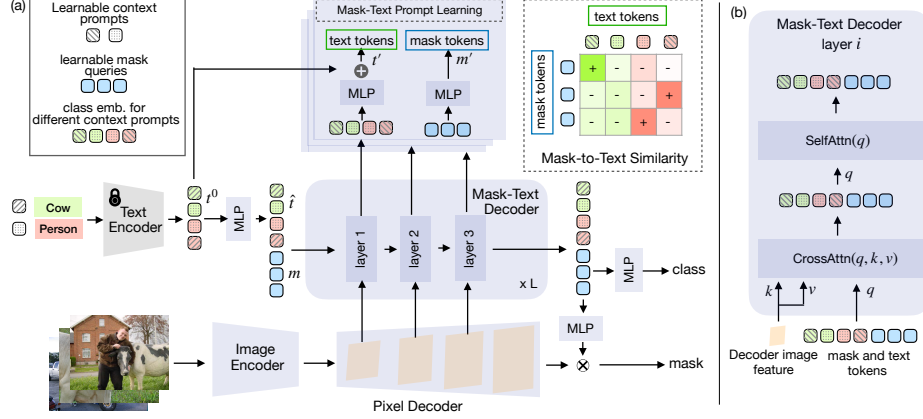
Fig. 3: Overview of our framework. (a) MTA-CLIP introduces Mask-Text Alignment (Sec. 3.2) that adopts a novel set of input queries combining both the mask queries (**m**) and text embeddings (**t̂**). These combined queries interact with each other within mask-text decoder layers, enhancing mask representations through the incorporation of rich textual data following **Text-Enhanced Mask Feature Learning**. Subsequently, class-level alignment is conducted between the mask and text representations following **Mask-to-Text Contrastive Learning** via a contrastive loss applied to the mask-to-text similarity scores. (b) Furthermore, distinct alignment of different context prompts with mask representations is achieved using **Mask-Text Prompt Learning**.

language model. Building upon this, CoCoOp [50] further refines the approach by incorporating image conditioning. Additionally, while the majority of existing methodologies concentrate on learning a singular prompt for improved generalization, ProDA [28] advocates for the utilization of multiple probabilistic prompts designed to cater to diverse contexts. In the context of semantic segmentation, DenseCLIP [31] adopts a CoCoOp [50] style of image-conditioned prompt learning for pixel-level predictions. Similarly, PPL [20] extends the ProDA framework [28] by learning a probabilistic distribution of diverse prompts tailored to different image contexts. Both DenseCLIP and PPL employ pixel-text alignment to facilitate prompt learning. In contrast, our approach involves Mask-Text Alignment for the acquisition of distinct prompts catering to diverse contexts.

3 MTA-CLIP

In this section, we first describe the preliminaries (Sec. 3.1), where we discuss mask-based segmentation method, Mask2Former [5] and Prompt Learning, as prior information required for our work. We then discuss the key component Mask-Text Alignment (Sec. 3.2), where we describe the two components, Mask-Text Decoder and Mask-Text Prompting, that effectively aligns the mask representations with the text embeddings. In the end, we present the training details (Sec. 3.3) for our framework.

3.1   Preliminaries

Mask2Former. Mask2Former [5] is a mask-based segmentation method that segments an image into N binary masks and classify them to K classes. Different from standard pixel segmentation methods, it consist of an additional lightweight transformer decoder along with the segmentation backbone and a pixel decoder. The transformer decoder takes learnable tokens, $m_1, \ldots, m_N$ (also called mask tokens) and pixel decoder features as input and processes these mask tokens with pixel decoder features with masked attention layers. It utilises pyramid resolution features extracted from the initial three layers of pixel decoder for processing mask tokens, whereas use the last layer (fourth layer) for making predictions. Typically, the dot product of the mask tokens with the last pixel decoder layer gives the final binary mask segmentations, whereas, an additional classification MLP predicts the class for these masks. Differently, in this work, our Mask-Text Decoder jointly process the text embedding and mask tokens. In particular, it enhances and align the mask tokens with CLIP text embeddings in different Mask-Text Decoder layers. Further, we also utilize multiple context prompts that captures diverse class representations among different masks.

Prompt Learning. CLIP [30] trained on 400M text-image pairs, learns powerful representations by aligning the image and text representations using contrastive learning. For better transferability on downstream tasks, prior works [20, 31, 51] learn contextual text prompts $\mathbf{p}$, instead of using hand crafted prompts such as "a photo of", or "a painting of". The prompts $\mathbf{p}$ are appended to the class token, $[\mathbf{p}, \texttt{CLS}]$ and are fed to the frozen text encoder to obtain the context rich class embeddings $\mathbf{t} = [t_1, \ldots, t_C]$ for C classes. Different from [20, 31], we learn prompts that align text features with mask representations. Moreover, we learn multiple context adaptive prompts to capture diverse class representations across different masks.

3.2   Mask-Text Alignment

The key component of MTA-CLIP is the novel Mask-Text Alignment mechanism, that combines class-specific language information from the CLIP text encoder with the mask representations. This alignment is achieved through two pivotal components: **(1) Mask-Text Decoder**, which effectively enriches and aligns mask representations with contextual language features to achieve more accurate and detailed segmentation, and **(2) Mask-Text Prompt Learning**, which utilizes multiple context-specific prompts to capture the diverse representations of mask segments, thereby addressing the variability in real-world scenarios.

1. Mask-Text Decoder. Our Mask-Text Decoder has two key components: 1) Text-Enhanced Mask Feature Learning, where we use the rich textual data from language models to enhance mask representations, creating more context-aware segmentation; and 2) Mask-to-Text Contrastive Learning, aiming to establish a

class-specific alignment between the mask representations and text embeddings within the CLIP feature space using a contrastive learning strategy.

Text-Enhanced Mask Feature Learning. In Text-Enhanced Mask Feature Learning, we enhance mask queries with class-specific language information derived from the CLIP text encoder. The process begins with constructing a novel set of queries $\mathbf{q}$, by concatenating mask queries $\mathbf{m} = [m_1, \ldots, m_N]$ with text embeddings $\mathbf{t} = [t_1, \ldots, t_C]$ for class labels from the CLIP text encoder, where N denotes the number of queries and C denotes the number of classes. Before concatenation, the text embeddings undergo dimension reduction through a MultiLayer Perceptron (MLP), adapting them to be compatible with the mask tokens in terms of dimension size and thus enhances computational efficiency. This reduction is formulated as:

$$\hat{\mathbf{t}} = \mathtt{MLP}(\mathbf{t}), \quad \mathbf{t} \in \mathbb{R}^{C \times 1024} \quad \text{and} \quad \hat{\mathbf{t}} \in \mathbb{R}^{C \times 256}. \tag{1}$$

Consequently, the resulting concatenated queries $\mathbf{q}$ are expressed as:

$$\mathbf{q} = [\hat{\mathbf{t}}, \mathbf{m}], \quad \mathbf{q} \in \mathbb{R}^{(C+N) \times 256}. \tag{2}$$

Through this concatenation, $\mathbf{q}$ integrates both CLIP-based image features and text embeddings, serving as the input of the Mask-Text Decoder.

Within the Mask-Text Decoder, which consists of L layers, $\mathbf{q}$ undergoes a series of layer-wise attention refinements. At each layer $i \in \{0, 1, \cdots, L-1\}$, the process includes cross-attention and self-attention as follows:

$$\mathbf{q}^{(i)}_{\mathrm{cross}} = \mathtt{CrossAttn}(\mathbf{q}^{(i)}, \mathbf{k}^{(i)}, \mathbf{v}^{(i)}), \tag{3}$$

$$\mathbf{q}^{(i+1)} = \mathtt{SelfAttn}(\mathbf{q}^{(i)}_{\mathrm{cross}}) \tag{4}$$

Here, $\mathbf{k}^{(i)}, \mathbf{v}^{(i)}$ are the ith level image features from the pixel decoder. During the cross-attention phase, both mask and text queries interact with image features, adapting to the specific context within the specific image. In the self-attention phase, mask queries not only interact among themselves to enhance the understanding of the relationships and context between different image areas, but also interact with text queries to integrate visual information with corresponding textual information.

Mask-to-Text Contrastive Learning. Building upon the Text-Enhanced Mask Feature Learning phase, where mask queries are augmented with visual and textual information, Mask-to-Text Contrastive Learning aims to establish class-level alignment between the mask and text representations within the CLIP feature space. This alignment is achieved through a contrastive learning approach, which involves enhancing the mask-text similarity score of a "positive" pair against those of "negative" pairs.

We first project $\mathbf{q}$, including both mask queries $\mathbf{m}$ and the text queries $\hat{\mathbf{t}}$, back to the CLIP feature space after each Transformer decoder layer. This projection is conducted using MultiLayer Perceptrons (MLPs):

$$\mathbf{t}' = \mathtt{MLP}(\hat{\mathbf{t}}) + \mathbf{t}^0, \quad \mathbf{m}' = \mathtt{MLP}(\mathbf{m}) \tag{5}$$

Here, the addition of the initial CLIP text embedding $\mathbf{t}^0$ to the projected text embeddings ensures the preservation of the original textual space from the CLIP model.

The essence of contrastive learning involves calculating the pairwise similarity between each projected mask query $m'_i, (i \in \{1, \cdots, N\})$ and text query $t'_j, (j \in \{1, \cdots, C\})$. The similarity score S, a dot product normalized by a temperature parameter τ, is given by:

$$S(m'_i, t'_j) = \frac{m'_i \cdot t'_j}{\tau}. \tag{6}$$

We select the text query with ground truth class for each mask query as the "positive" class, while "negative" pairs include the mask query and all other non-corresponding text queries. The alignment is refined through a contrastive loss function, $\mathcal{L}_{sim}$, which is designed to maximize the similarity for the positive pair and minimize it for negative pairs. The loss is defined as:

$$\mathcal{L}_{sim} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \left[y_{ij} \cdot \log \frac{\exp\left(S(m'_i, t'_j)\right)}{\exp\left(S(m'_i, t'_j)\right) + \sum_{\substack{r \in \{1, \cdots, C\} \\ r \neq j}} \exp\left(S(m'_i, t'_r)\right)} \right] \tag{7}$$

Here, y_{ij} equals 1 for the mask-text pair where the text query belongs to the ground truth class for the given mask query, and 0 otherwise. We obtain the ground truth class for the given mask query using bipartite matching of the predicted binary mask with actual ground truth masks similar to [5]. This contrastive learning process aligns each mask query with its corresponding class text query. Moreover, we apply $\mathcal{L}_{sim}$ to $\mathbf{m}'$ and $\mathbf{t}'$ after each Mask-Text Decoder layer utilizing hierarchical feature pyramids from pixel decoder, enabling the model to accurately capture and segment across a range of mask sizes.

2. Mask-Text Prompt Learning. Our approach tailors text prompts to distinctively align with the diverse contexts of mask segments within the same class. By implementing K unique prompts shared across C classes, we create an extensive range of $K \times C$ text queries from the text encoder. This expanded set of text queries is specifically designed to cater to the varied contexts in which a class can appear. By aligning each mask segment with this extensive array of class-context combinations, our model achieves a fine-grained Mask-Text alignment, leading to more precise and context-aware masks. To accommodate this variety in contextual representations, we expand the query set in Text-Enhanced Mask Feature Learning from $C + N$ to $K \times C + N$. For Mask-to-Text Contrastive Learning, we adopt a refined approach: For each mask token m'_i, the text query among the different K instances from its ground truth class with the highest similarity score is identified as the "positive" query. Specifically, the positive text query for mask token m'_i with ground truth class j is defined as : $t'_{j,\widetilde{k}}$, such that $\widetilde{k} = \arg\max_{k \in \{1, \cdots, K\}} S(m'_i, t'_{j,k})$. We employ two distinct strategies for selecting negative pairs:

MixNeg: All other text queries except the positive one (totaling $K \times C - 1$) are treated as negative, with the loss modified as:

$$\mathcal{L}_{sim} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \left[y_{ij} \cdot \log \frac{\exp\left(S(m_i', t_{j,\tilde{k}}')\right)}{\exp\left(S(m_i', t_{j,\tilde{k}}')\right) + \sum_{\substack{r \in \{1, \cdots C\}, k \in \{1, \cdots K\} \\ \{r,k\} \neq \{j,\tilde{k}\}}} \exp\left(S(m_i', t_{r,k}')\right)} \right] \quad (8)$$

SeperateNeg: Text queries not belonging to the ground truth class (totaling $K \times (C-1)$) are treated as negative, with the loss modified as:

$$\mathcal{L}_{sim} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \left[y_{ij} \cdot \log \frac{\exp\left(S(m_i', t_{j,\tilde{k}}')\right)}{\exp\left(S(m_i', t_{j,\tilde{k}}')\right) + \sum_{\substack{r \in \{1, \cdots C\}, k \in \{1, \cdots K\} \\ r \neq j}} \exp\left(S(m_i', t_{r,k}')\right)} \right] \quad (9)$$

As detailed in Sec. 4.2, both strategies exhibit improvement, with **_MixNeg_** emerging as more effective due to its superior capability in differentiating between closely related contexts.

3.3 Training and Inference

Our training process combines the strengths of both vision and language models to optimize the semantic segmentation performance of MTA-CLIP. The core of our training lies in the effective integration and alignment of mask representations with text embeddings through the Mask-Text Decoder.

Multi-Layer Mask-Text Alignment: At each layer of the Mask-Text Decoder, we handle N mask tokens alongside $K \times C$ class tokens derived from our Mask-Text Prompt Learning. The height and width of the pixel decoder image feature used at each Mask-Text Decoder layer, denoted as $h^{(i)}$ and $w^{(i)}$, vary depending on the layer. Similar to [5], these dimensions are set as $(h^{(i)}, w^{(i)}) = (H/32, W/32), (H/16, W/16), (H/8, W/8)$ for layers $i \bmod 3 = 0, 1, 2$, respectively, where H and W are the height and width of the original image, and $i \in [0, \cdots, 8]$ for Mask-Text Decoder with 9 layers. This variation in feature resolution across layers ensures Mask-Text Alignment at different scales, important for segmenting classes of different sizes.

Overall Loss Function: We adopt a composite loss function that balances mask classification, mask segmentation, and Mask-Text Alignment. Specifically, we use cross-entropy loss $\mathcal{L}_{mask-cls}$ for the classification prediction of mask tokens and binary cross-entropy loss $\mathcal{L}_{mask}$ for the segmentation prediction following [5]. Additionally, we have the similarity loss $\mathcal{L}_{sim}$ from Equation 8 for aligning mask tokens with context-specific class tokens. The final loss function is a weighted sum of these components:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{mask-cls} + \lambda_2 \mathcal{L}_{mask} + \lambda_3 \mathcal{L}_{sim}, \quad (10)$$

where λ_1, λ_2, and λ_3 are weighting factors balancing the contribution of each loss component to the overall training objective.

Inference: During inference, we ignore the text tokens $(t_1, ..., t_C)$ and utilize only the mask tokens $(m_1, ..., m_N)$ from the final layer (9th layer) of the Mask-Text Decoder for final mask predictions. In particular, following [5] the text

aligned mask tokens $(m_1, ..., m_N)$ are fed to an MLP layer, followed by dot product with pixel-decoder final layer features to give N masks for the mask tokens. Moreover, these text aligned masks tokens goes through a separate classification MLP layer to obtain the class prediction for these mask tokens.

4 Experiments

In this section, we discuss the extensive experiments we conducted to demonstrate the effectiveness of MTA-CLIP. We compare our framework's performance with state-of-the-art methods for semantic segmentation. Then, we discuss ablation experiments followed by qualitative results.

Datasets and Evaluation Metric. We evaluate MTA-CLIP on two standard dataset benchmarks i.e., ADE20k [48] and Cityscapes [8]. ADE20k has 150 categories, covering both indoor and outdoor scenes. It has around 20,000 images for training and 2,000 images for validation. Cityscapes, on the other hand, is an urban scene understanding dataset comprising 19 categories. It has 2,975 training images along with 500 validation images. Following prior works [5,20,31], we report the mean Intersection over Union (mIoU) score for evaluation.

Implementation Details. We employ Mask2Former [5] as our baseline segmentation network with CLIP pre-trained backbone for the image encoder. Following previous works [20,31], we leverage all three available CLIP image backbones: ResNet-50 [15], ResNet-101 [15], and ViT-B [12]. The text encoder is based on the CLIP-trained model, and we maintain its frozen state during training. In line with [31], we set the context length for each context prompt $\mathbf{p}$ to 8. Moreover, we conduct experiments with $K = 3$ different contextual prompts, unless otherwise specified. Following prior works in the field [5,6], we set $N = 100$ as the number of learnable mask tokens. Our final loss ($\mathcal{L}$) incorporates parameters $\lambda_1 = 2.0$, $\lambda_2 = 5.0$, and $\lambda_3 = 1.0$. Training hyperparameters, following [31], are as follows: for the ADE20k dataset, we utilize a crop size of 512×512, a batch size of 16, and train for 160k iterations. On the Cityscapes dataset, a crop size of 1024×1024, a batch size of 8, and 90k iterations are employed. The AdamW optimizer [27] is used with an initial learning rate of 2×10^{-4}.

4.1   Comparison with State-of-the-Art

We compare MTA-CLIP's performance on two standard benchmarks i.e., ADE20k in Tab. 1 and Cityscapes in Tab. 2 with different backbones. We make two key comparisons. First, we compare our framework's performance with prior state-of-the-art's that uses VLM such as CLIP for semantic segmentation task. These methods [20,31] employ pixel-text alignment for aligning the vision and language modalities. Second, we compare the performance of our frameworks with an intuitive baseline of Mask2Former [5] with CLIP trained image encoder, without any Mask-Text Alignment. We also compare MTA-CLIP performance with classical semantic segmentation methods with ImageNet pre-training. Due to space constraint, we show Params and FLOPs comparison in supplement (Tab. 4).

Table 1: Semantic Segmentation results on ADE20k dataset. We show comparisons of our method with prior works for different backbones. The mIoU (SS) and mIoU (MS) are single-scale and multi-scale inference results, respectively. *: We re-implement Mask2Former with CLIP pre-trained backbones.

Backbone	Method	Pre-train	mIoU (SS)	mIoU (MS)
ResNet-50	PSPNet [45]	ImageNet	41.1	41.9
	DeepLabv3+ [4]	ImageNet	42.7	43.8
	UperNet [38]	ImageNet	42.1	42.8
	Mask2Former [5]	ImageNet	47.2	49.2
	Mask DINO [23]	ImageNet	47.7	-
	Semantic FPN [18]	ImageNet	38.6	40.6
	CLIP + Semantic FPN [30]	CLIP	39.6	41.6
	DenseCLIP [31]	CLIP	43.5	44.7
	PPL [20]	CLIP	44.7	45.8
	Mask2Former* [5]	CLIP	47.4	49.0
	Ours	CLIP	**49.1**	**49.9**
ResNet-101	PSPNet [45]	ImageNet	43.6	44.4
	DeepLabv3+ [4]	ImageNet	44.6	46.1
	UperNet [38]	ImageNet	43.8	44.8
	Mask2Former [5]	ImageNet	47.7	49.6
	Semantic FPN [18]	ImageNet	40.4	42.3
	CLIP + Semantic FPN [30]	CLIP	42.7	44.3
	DenseCLIP [31]	CLIP	45.1	46.5
	PPL [20]	CLIP	46.4	47.8
	Mask2Former* [5]	CLIP	48.4	49.3
	Ours	CLIP	**49.8**	**51.1**
ViT-B	SETR-MLA-DeiT [46]	ImageNet	46.2	47.7
	Mask2Former [5]	ImageNet	51.1	51.9
	BEiT [37]	ImageNet	47.7	-
	Semantic FPN [18]	ImageNet-21K	49.1	50.9
	CLIP + Semantic FPN [30]	CLIP	49.4	50.3
	DenseCLIP [31]	CLIP	50.6	51.3
	PPL [20]	CLIP	51.6	51.8
	Mask2Former* [5]	CLIP	51.0	51.9
	Ours	CLIP	**52.3**	**53.1**

Results for ADE20k. MTA-CLIP exhibits superior performance over the state-of-the-art Pixel-to-Text alignment segmentation method PPL [20], achieving a performance gain of 4.4% and 3.4% for ResNet-50 and ResNet-101 backbones, respectively. This emphasizes the substantial impact of our Mask-Text Alignment approach compared to the Pixel-Text alignment approach for semantic segmentation tasks. Employing the ViT-B backbone, our framework achieves a performance of 52.3%, surpassing PPL's 51.6%. Furthermore, our method outperforms the prior state-of-the-art ImageNet pretrained Mask2Former by 1.9%, 2.1%, and 1.2% for ResNet-50, ResNet-101, and ViT-B backbones, respectively.

Results for Cityscapes. MTA-CLIP exhibits superior performance compared to the Pixel-to-Text alignment-based segmentation method, DenseCLIP [31], with gains of 2.2% and 1.1% for ResNet-50 and ResNet-101 backbones, respectively. Mask2Former with CLIP pretraining with our reimplementation improves over Mask2Former with ImageNet pretraining by 0.6% and 1% for ResNet-50 and ResNet-101 backbones, respectively. Our framework further surpasses this

Table 2: Semantic Segmentation results on Cityscapes dataset. We show comparisons of our method with two types of pretraining of image encoder (ImageNet and CLIP) for different backbones. * : We reimplement Mask2Former with CLIP pretrained backbone.

ResNet-50				ResNet-101			
Method	Pre-train	mIoU (SS)	mIoU (MS)	Method	Pre-train	mIoU (SS)	mIoU (MS)
PSPNet [45]	ImageNet	78.5	79.8	PSPNet [45]	ImageNet	79.8	81.0
DeepLabv3+ [4]	ImageNet	80.1	81.1	DeepLabv3+ [4]	ImageNet	81.0	82.0
Mask2Former [5]	ImageNet	79.4	82.2	Mask2Former [5]	ImageNet	80.1	81.9
UperNet [38]	ImageNet	78.2	79.2	UperNet [38]	ImageNet	79.4	80.4
DenseCLIP [31]	CLIP	79.2	80.3	DenseCLIP [31]	CLIP	81.0	81.8
Mask2Former* [5]	CLIP	81.0	82.1	Mask2Former* [5]	CLIP	81.1	82.4
Ours	CLIP	**81.7**	**82.8**	Ours	CLIP	**82.1**	**83.6**

strong baseline by 0.7% and 1.0% for ResNet-50 and ResNet-101 backbones, respectively, showcasing the importance of Mask-Text alignment.

In summary, MTA-CLIP consistently outperforms previous methodologies in semantic segmentation by a substantial margin across standard dataset benchmarks. This observation shows the enhanced performance achieved through additional language alignment for masks. Additionally, our framework also exhibits superiority over previous semantic segmentation methods with ImageNet pre-training, emphasizing the significance of leveraging large-scale Vision-Language Models (VLMs) for network backbone pre-training.

4.2   Ablation Analysis

Ablation with different components. We conduct ablation studies on the components of our network, as detailed in Tab. 3. 'Baseline' denotes Mask2Former trained with the CLIP ResNet-50 backbone. We first enhance mask representations following Text-Enhanced Mask Feature Learning (Text-Enhanced, Sec. 3.2) resulting in a performance gain of 0.6% mIoU over baseline. Subsequently, by aligning mask representations with text representations with similarity loss using Mask-to-Text Contrastive Learning (MTCL, Sec. 3.2), we observe a further increment of 0.5% in mIoU. Finally, we use Mask-Text Prompt Learning (MTPL, Sec. 3.2) with *MixNeg* to learn distinct prompts, yielding an additional boost of 0.5% in mIoU. The Mask-Text Decoder, which combines the learning of enhanced mask representations (Text-Enhanced) followed by aligning them with text representations (MTCL), gives a significant boost of 1.2% in the mIoU score compared to the baseline, showing the importance of such alignment.

Results with other backbones. We extend our evaluation by comparing the performance of MTA-CLIP with alternative backbones, as presented in Tab. 3. Specifically, we utilize ImageNet [9] pretrained ViT-S [30] and MiT-B4 [39] backbones for this experiment. We demonstrate the versatility of our framework, indicating its applicability beyond Vision-Language Model (VLM)-trained backbones to also include ImageNet trained alternatives. In the case of ViT-S, Dense-CLIP [31] based on pixel-text alignment achieves of 47.8% mIoU. Our MTA-CLIP framework, relying on Mask-Text Alignment, surpasses DenseCLIP by

3.6% in mIoU and also outperforms the strong baseline Mask2Former by 0.8%. Moreover, our framework exhibits effective compatibility with Mix Transformer backbones [39], demonstrating a performance improvement of 1.4% compared to SegFormer [39], which employs a similar backbone. Collectively, these results affirm the adaptability of our framework across diverse backbones.

Table 3: Left: Semantic Segmentation result with other backbones (ViT-S and MiT-B4). The experiments are performed on ADE20k dataset. Right: Ablation with different components of our method. Initially, we refine the mask representations using Text-Enhanced Mask Feature Learning (Text-Enhanced, Sec. 3.2), followed by aligning the mask with text representations with Mask-to-Text Contrastive Learning (MTCL, Sec. 3.2). Subsequently, we learn distinct context prompts using Mask-Text Prompt Learning (MTPL Sec. 3.2). The term 'Baseline' refers to Mask2Former [5].

Backbone	Method	mIoU
ViT-S [12]	Segmenter [33]	46.2
	DenseCLIP [31]	47.8
	Mask2Former [5]	50.6
	Ours	**51.4**
MiT-B4 [39]	SegFormer [39]	50.3
	SegFormer (mask decoder)	51.2
	Ours	**51.7**

Methods	mIoU
Baseline	47.4
+ Text-Enhanced	48.1
+ Text-Enhanced + MTCL	48.6
+ Text-Enhanced + MTCL + MTPL	**49.1**

Enhancing Mask with Textual Data. In this analysis (Tab. 4), we underscore the significance of enhancing mask representations with textual data from the pre-trained CLIP language model. Specifically, we modify the SelfAttn operation outlined in Eq. (4) to ensure that the text embeddings $\hat{t}$ and mask queries m do not attend to each other. Instead, individual SelfAttn operations are applied to the text embeddings $\hat{t}$ and mask queries m, denoted as $\hat{t} = \text{SelfAttn}(\hat{t})$ and $m = \text{SelfAttn}(m)$. It is important to note that, in this experiment, only the SelfAttn operation is modified, while all other operations remain the same. The results reveal a performance decline of 0.5% when mask queries m are not attended by the text embeddings $\hat{t}$. This observation shows the importance of attending mask queries with text embeddings for optimal performance.

Comparison between Contextual Prompts. As discussed in Sec. 3.2, we employ two distinct strategies, namely ***SeparateNeg*** and ***MixNeg***, to derive specialized text prompts aligned with the diverse contexts of masks for a given class. In this experimental analysis, we first compare the efficacy of the two strategies, *SeparateNeg* and *MixNeg*. Subsequently, we assess our framework's performance across different numbers of text prompts (K). Overall, both approaches show effectiveness (see Tab. 4), with *MixNeg* exhibiting a performance gain of 0.5%, compared to 0.3% achieved by *SeparateNeg*, over the baseline that utilizes a single prompt. Furthermore, our observations indicate that employing $K = 3$, involving three learnable unique prompts shared across all classes, yields the optimal performance of 49.1% with *MixNeg*. Similarly, for *SeparateNeg*, the best performance is achieved with $K = 4$ across varying values of K.

Table 4: Left: Importance of enhancing mask with rich text data. Experiments on ADE20k dataset. Right: Ablation with contextual prompts. In this experiment, we compare the two different approaches for effectively using multiple prompts as described in.... '# prompts' refers to the number of different prompts used as input to CLIP text encoder.

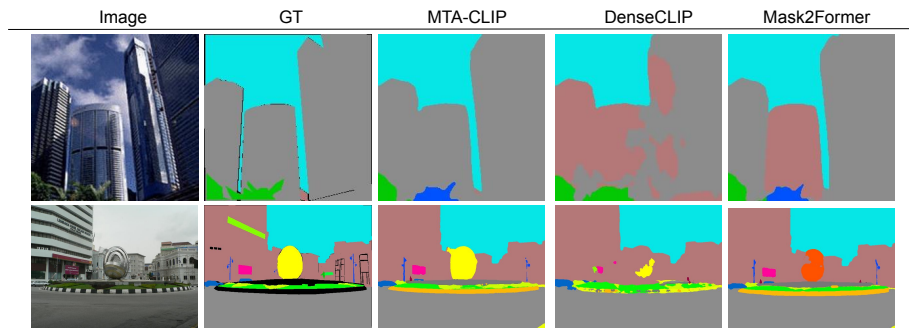|			# prompts	2	3	4	5
Baseline		47.4	SeparateNeg	48.5	48.7	**48.9**	48.2
+ Separate attentions in text and mask queries		48.6	MixNeg	48.6	**49.1**	48.9	48.4
+ Joint attention in text and mask queries		49.1					

Fig. 4: Qualitative comparison. We compare our framework, MTA-CLIP's performance with Pixel-Text Alignment segmentation approach, DenseCLIP [31] and Mask2Former [5] baseline with ResNet-50 backbone.

4.3   Qualitative Analysis

We present the qualitative comparison of our framework in Fig. 4. We compare our framework with Pixel-to-Text Alignment method DenseCLIP [31] and mask-based semantic segmentation baseline Mask2Former [5] for ResNet-50 backbone on ADE20k dataset. Overall, our framework segments the classes better than DenseCLIP and Mask2Former. Additional results with ResNet-101 backbone and on Cityscapes dataset is in supplement (Sec. 2).

5   Conclusion

This work introduces a novel framework, MTA-CLIP, designed for the semantic segmentation task, which implements innovative mask-level vision-language alignment. The proposed framework enhances mask representations by incorporating and aligning them with class-specific language features. Additionally, multiple context-specific prompts are employed to capture diverse representations of the mask segments. The framework is systematically evaluated using various backbones and pretraining strategies, consistently surpassing prior state-of-the-art approaches across diverse dataset benchmarks in the semantic segmentation task. We hope that this work stimulates further innovation in this challenging yet promising research direction.

References

1. Bertasius, G., Shi, J., Torresani, L.: Semantic segmentation with boundary neural fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3602–3610 (2016)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2018). https://doi.org/10.1109/TPAMI.2017.2699184
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
7. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4113–4123 (2024)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6819–6829 (2019)
11. Ding, J., Xue, N., Xia, G., Dai, D.: Decoupling zero-shot semantic segmentation. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11573–11582 (2021)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
13. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision pp. 1–15 (2023)
14. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision. pp. 540–557. Springer (2022)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

16. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)

17. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)

18. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6399–6408 (2019)

19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

20. Kwon, H., Song, T., Jeong, S., Kim, J., Jang, J., Sohn, K.: Probabilistic prompt learning for dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6768–6777 (2023)

21. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.243, https://aclanthology.org/2021.emnlp-main.243

22. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022)

23. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3041–3050 (2023)

24. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 510–519 (2019)

25. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.353, https://aclanthology.org/2021.acl-long.353

26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7

28. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5206–5215 (2022)

29. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4353–4361 (2017)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
31. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
33. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
34. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5229–5238 (2019)
35. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European conference on computer vision. pp. 108–126. Springer (2020)
36. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. TPAMI (2019)
37. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
38. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
39. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
40. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
41. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
42. Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N.: Context prior for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12416–12425 (2020)
43. Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J.: Lite-hrnet: A lightweight high-resolution network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10440–10450 (2021)

44. Yuan, Y., Xie, J., Chen, X., Wang, J.: Segfix: Model-agnostic boundary refinement for segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 489–506. Springer (2020)
45. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
46. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
47. Zheng Ding, Jieke Wang, Z.T.: Open-vocabulary universal image segmentation with maskclip. In: International Conference on Machine Learning (2023)
48. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
49. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
50. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
51. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)