

ITACLIP: Boosting Training-Free Semantic Segmentation with Image, Text, and Architectural Enhancements

M. Arda Aydın
 Bilkent University
 arda.aydin@bilkent.edu.tr

Efe Mert Çırpar
 RWTH Aachen University
 efe.cirpar@rwth-aachen.de

Elvin Abdinli
 Technical University of Munich
 elvin.abdinli@tum.de

Gozde Unal
 Istanbul Technical University
 gozde.unal@itu.edu.tr

Yusuf H. Sahin
 Istanbul Technical University
 sahinyu@itu.edu.tr

Abstract

Recent advances in foundational Vision Language Models (VLMs) have reshaped the evaluation paradigm in computer vision tasks. These foundational models, especially CLIP, have accelerated research in open-vocabulary computer vision tasks, including Open-Vocabulary Semantic Segmentation (OVSS). Although the initial results are promising, the dense prediction capabilities of VLMs still require further improvement. In this study, we enhance the semantic segmentation performance of CLIP by introducing new modules and modifications: 1) architectural changes in the last layer of ViT and the incorporation of attention maps from the middle layers with the last layer, 2) Image Engineering: applying data augmentations to enrich input image representations, and 3) using Large Language Models (LLMs) to generate definitions and synonyms for each class name to leverage CLIP’s open-vocabulary capabilities. Our training-free method, ITACLIP, outperforms current state-of-the-art approaches on five popular segmentation benchmarks. Our code is available at <https://github.com/m-arda-aydn/ITACLIP>.

1. Introduction

The emergence of large foundational Vision Language Models (VLMs) [2, 29, 30, 49] has driven a paradigm shift in deep learning. Before the introduction of these foundational models, computer vision models were effective only for a limited number of classes. However, this evaluation setting does not accurately reflect real-world conditions, as the world is not limited to a small number of classes, and models trained on these finite classes exhibit weak transferability to real-world scenarios. These reasons highlight the need for open-vocabulary models rather than relying on

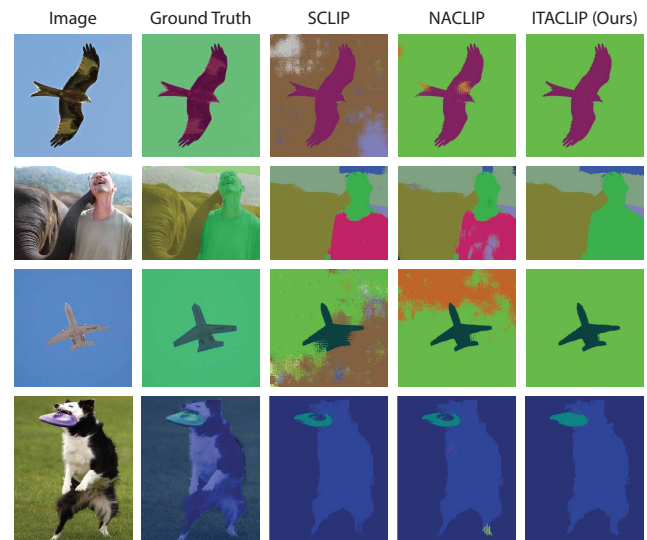


Figure 1. **Qualitative comparison of training-free semantic segmentation methods.** We compare ITACLIP with SCLIP [61] and NACLIP [25] using images from the COCO-Stuff [8] dataset. Additional visualizations are included in the Appendix.

predefined classes.

Rapid advancements in VLMs have fueled research on open-vocabulary computer vision tasks. In particular, the CLIP [49] model has become foundational for various computer vision tasks such as multi-label image classification [1, 42], object detection [18, 24, 55], semantic segmentation [5, 33, 36, 37, 41, 52, 54, 56, 57, 71], and image generation [59, 62]. The generalization capability of CLIP has enabled models to achieve remarkable results in open-vocabulary settings with relatively minimal modifications instead of requiring training from scratch.

Despite the great success of VLMs in zero-shot im-

age classification [30, 49], translating this achievement into dense prediction tasks has become challenging for researchers. Although several works [16, 68, 71] have demonstrated strong results in zero-shot semantic segmentation by employing new components, these models still need expensive *pixel-level* annotations of seen classes.

Segment Anything Model (SAM) [32], a large foundational segmentation model, is trained on a massive dataset and delivers impressive results in semantic segmentation. Yet, the segmentation capability of SAM strongly depends on the contents of its training dataset, and further studies [14, 28, 45] indicate that its performance degrades when applied to images dissimilar to those in the training data. Furthermore, collecting annotated data can be quite expensive in certain domains, such as biomedical image analysis, where labeling requires specialized expertise. This limitation significantly impacts the scalability of models. In response to these challenges, researchers have turned their focus to training-free semantic segmentation models as a promising solution. Training-free semantic segmentation models typically use a VLM like CLIP and aim to transform VLM’s image-level knowledge into pixel-level predictions for dense prediction tasks [5, 6, 25, 37, 58, 61, 63, 70].

In this study, we present **ITACLIP** (Image-Text-Architectural Enhanced **CLIP**), our training-free method, which utilizes architectural changes and enriches the input features to improve the segmentation performance. Our approach outperforms the current state-of-the-art methods on COCO-Stuff [8], COCO-Object [40], Pascal Context [47], Pascal VOC [20], and Cityscapes [12]. As illustrated in Fig. 1, ITACLIP produces more accurate segmentation maps than SCLIP [61] and NACLIP [25].

Our contributions can be summarized as follows:

- We employ a combination of *self-self attentions* instead of the original attention mechanism of ViT [17]. Moreover, we *remove* the Feed-Forward Network (FFN) in the last layer of ViT and combine the *attention maps* from the middle layers with the attention map of the final layer.
- We leverage Large Language Models (LLMs) to develop a systematic strategy for *generating auxiliary texts* for any class name in an open-vocabulary setting. We integrate text features from the original class names with those from definitions or synonyms.
- We propose the *Image Engineering* module, a novel component designed to refine features extracted from the image encoder. Rather than feeding only the original image to CLIP’s visual encoder, we also utilize augmented images to enrich and diversify the overall image representation.
- We present a comprehensive analysis demonstrating that ITACLIP achieves state-of-the-art results on various segmentation benchmarks.

2. Related Work

2.1. Foundational Vision-Language Models

Inspired by the success of large-scale foundational models in natural language processing [15, 48, 50, 51], researchers have turned their attention to training foundational models in computer vision. As part of this effort, VLMs aim to integrate textual and visual information to construct a unified understanding. Contrastive learning-based VLMs have demonstrated impressive capabilities in visual-language understanding [2, 3, 11, 29, 30, 35, 49, 64].

Contrastive Language-Image Pre-training (CLIP) [49] includes both an image encoder and a text encoder, jointly trained on a private WIT-400M dataset containing image-text pairs. CLIP has exhibited strong transferability in zero-shot visual recognition tasks, fundamentally changing the paradigm for zero-shot learning in computer vision. Like CLIP, ALIGN [30] employs a dual-encoder architecture; however, it is trained on a much larger private dataset to enhance scalability and simplify dataset curation. OpenCLIP [29] is an open-source implementation of CLIP and is trained on the publicly available LAION [53] dataset. Due to the strong transferability of CLIP on downstream tasks, many training-free semantic segmentation models [33, 61] employ OpenAI’s pre-trained CLIP model. Hence, our method also follows this implementation.

2.2. Open-Vocabulary Semantic Segmentation

Semantic segmentation can be defined as the pixel-wise classification of an image. Conventionally, semantic segmentation models are trained on a limited set of known classes, and the evaluation process is conducted for these specific classes. In contrast, open-vocabulary semantic segmentation models use VLMs to assign *arbitrary* class labels to each pixel in a given image. The target class set includes not only predefined classes but also arbitrary class names. Open-vocabulary semantic segmentation models can be categorized into three groups: 1) **Training-Free** Methods, 2) **Weakly Supervised** Methods, and 3) **Fully Supervised** Methods.

Training-Free Methods. Training-free semantic segmentation models make predictions without using pixel-level or image-level annotations. MaskCLIP [70] discards the query and key embeddings, using only the value embedding from the final self-attention block while modifying the last linear layer of CLIP’s visual encoder. Inspired by MaskCLIP, research on training-free semantic segmentation models has rapidly accelerated. SCLIP [61] introduces a new attention formula based on query-query and key-key attention, leading to improved performance compared to MaskCLIP. GEM [6] employs generalized self-self attention combinations along with certain regularization steps. NACLIP [25] incorporates a Gaussian window

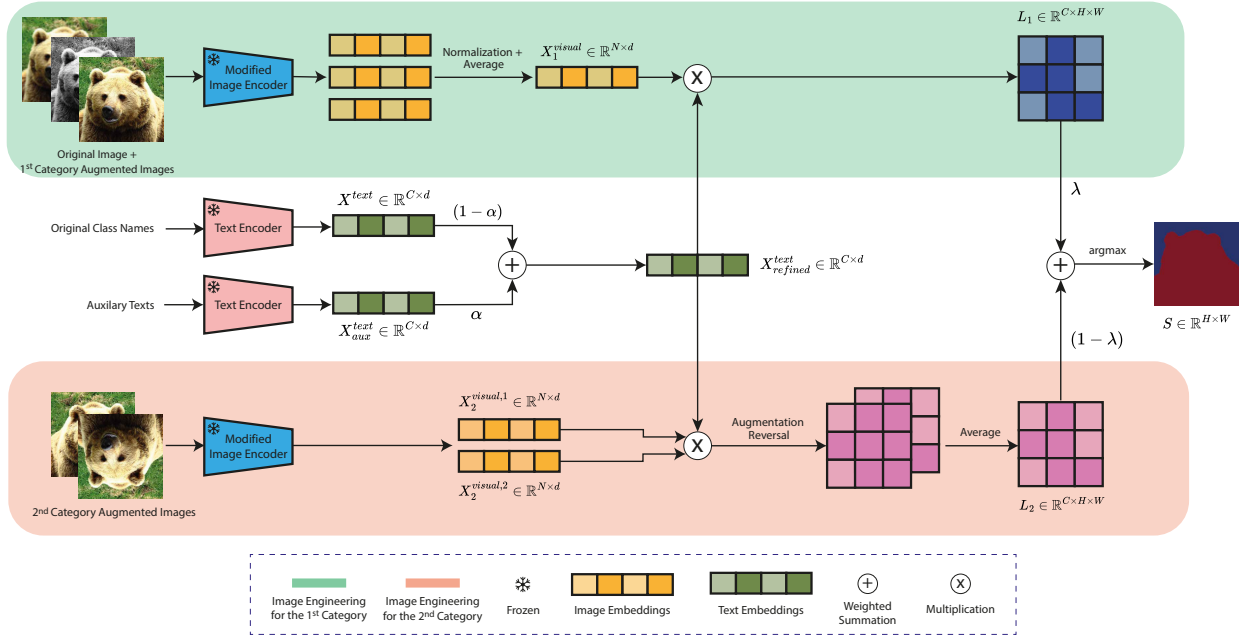


Figure 2. **Overview of ITACLIP.** Our method integrates image, text, and architectural enhancements to produce a more accurate segmentation map. We apply various data augmentation techniques, then process both the original and augmented images through a modified image encoder to obtain image embeddings. We also utilize an LLM to generate auxiliary texts (e.g., definitions or synonyms) for each original class name. The λ and α symbols denote the image engineering and auxiliary text coefficients used in weighted summations, respectively.

into the attention map of each patch to increase spatial attention around the patch and modifies CLIP’s visual encoder to boost the model’s performance. CLIP-DIY [63] extracts patch-level features from various image crops and aggregates them before upsampling. Alternatively, some methods [42, 58] utilize Class Activation Maps (CAMs) [69] to localize regions activated by the target class and then perform segmentation based on these activated areas. TagCLIP [42] first performs multi-label classification using modules designed for classification and feeds the predicted classes to CLIP-ES [41], a CAM-based segmentation framework, for segmentation. CaR [58] operates in two-stage units and recurrently generates masks until the predicted classes remain unchanged. The two-stage unit comprises two components: a mask generator (using CLIP-ES) and a mask classifier (using CLIP).

Weakly Supervised Methods. Weakly supervised semantic segmentation (WSSS) models [10, 41, 44, 52, 65, 66] rely on weak supervision, such as image-level labels, rather than labor-intensive pixel-level annotations. GroupViT [65] learns visual representations by grouping semantically related regions within images. TCL [10] leverages text-grounded images and introduces a “text-grounded contrastive loss” to align captions with corresponding regions. OVSegmentor [66] uses Slot Attention [43] to group visual features and aligns these groups with related text fea-

tures.

Fully Supervised Methods. Fully supervised semantic segmentation models [23, 26, 31, 34, 38, 39, 67] are typically trained on a specified dataset with pixel-level annotations available. Since these models have access to fine-grained dense labels, they generally outperform training-free and weakly supervised models.

Our study focuses on training-free semantic segmentation, aiming to extend CLIP’s image-level capabilities to the pixel-level without relying on additional training or external models trained on *pixel-level* annotations, like SAM. Thus, ITACLIP does not require pixel-level annotations at any stage to produce segmentation maps.

3. Method

In this section, we introduce our training-free semantic segmentation model, ITACLIP, as depicted in Fig. 2. First, we revisit the original ViT-based CLIP image encoder¹. Next, we begin our investigation by examining a vanilla approach—a straightforward method utilizing patch tokens for segmentation—as our baseline study. Finally, we present the architectural modifications to the CLIP’s ViT

¹Note that we focus exclusively on ViT-based [17] image encoders due to the limited zero-shot capabilities of ResNet-based [27] encoders in dense prediction tasks, as demonstrated in [70].

and introduce the proposed modules designed to expand the input representation.

3.1. Revisiting Original CLIP

The original ViT-based CLIP [49] consists of two distinct transformer-based encoders: an image encoder and a text encoder. The image encoder takes an input image $I \in \mathbb{R}^{3 \times H \times W}$ and divides it into non-overlapping patches of size $P \times P$, where H and W represent the height and width of the image, respectively. Thus, we obtain N different patch tokens, where $N = HW/P^2$ is the total number of patches. Subsequently, a linear projection maps each patch token to a d -dimensional space, with d denoting the feature dimension of the model. Additionally, the class token, $[\text{CLS}]$, is concatenated with the patch tokens extracted from image patches to capture the global context of the image. Lastly, positional embeddings are added to the visual tokens to prepare them as input for the first layer of the encoder.

The l th layer of the visual encoder receives visual tokens $X^{(l-1)} = [x_{\text{CLS}}, x_1, \dots, x_N] \in \mathbb{R}^{(N+1) \times d}$ from the previous layer. Next, $X^{(l)}$ is calculated as,

$$X^* = X^{(l-1)} + \mathbf{SA}(\text{LN}(X^{(l-1)})) \quad (1)$$

$$X^{(l)} = X^* + \mathbf{MLP}(\text{LN}(X^*)) \quad (2)$$

where LN , \mathbf{SA} , and \mathbf{MLP} represent Layer Normalization, the Self-Attention module, and the Feed-Forward Network (FFN), respectively.

Given an input $X \in \mathbb{R}^{(N+1) \times d}$, the Self-Attention module can be mathematically described as,

$$\mathbf{Attn}(X) = \text{softmax}\left(\frac{XW_QW_K^TX^T}{\sqrt{d}}\right) \quad (3)$$

$$\mathbf{SA}(X) = \mathbf{Attn}(X)XW_VW_O \quad (4)$$

where $\mathbf{Attn}(X) \in \mathbb{R}^{(N+1) \times (N+1)}$ denotes the attention map for input X , and $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d \times d}$ represent query, key, value, and output projection matrices learned during pre-training, respectively. For simplicity, we consider only single-head self-attention in our self-attention module description.

3.2. Vanilla Approach

Applying a patch-based classification strategy is the most straightforward approach to utilize CLIP for segmentation. We employ CLIP's image encoder to generate visual features $X^{\text{visual}} = [x_{\text{CLS}}^{\text{visual}}, X_{\text{patch}}^{\text{visual}}] \in \mathbb{R}^{(N+1) \times d}$ of given image I , as described in Sec. 3.1. $x_{\text{CLS}}^{\text{visual}} \in \mathbb{R}^{1 \times d}$ and $X_{\text{patch}}^{\text{visual}} \in \mathbb{R}^{N \times d}$ denote the extracted features from the $[\text{CLS}]$ token and image patches, respectively. For patch-based classification, we use only the features obtained from

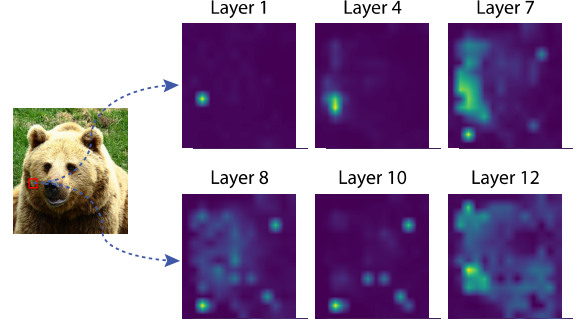


Figure 3. **Visualization of attention maps from various layers for a selected patch.** The red rectangle indicates the position of the randomly selected patch. Note that we use CLIP-ViT-B/16 as our visual backbone, with Layer 12 serving as the final layer.

patches, $X_{\text{patch}}^{\text{visual}}$. CLIP's text encoder is leveraged to compute the textual embeddings, $X^{\text{text}} \in \mathbb{R}^{C \times d}$, for the target C classes. Then, we use cosine similarity to determine similarity scores between the text embeddings and patch features. After the softmax operation, we interpolate the resulting logit to resize it to the original image dimensions. Finally, we perform a class-wise argmax operation to produce the segmentation map $S \in \mathbb{R}^{H \times W}$. Formally,

$$S = \underset{c \in C}{\text{argmax}}(\text{upsample}(\cos(X_{\text{patch}}^{\text{visual}}, X^{\text{text}}))) \quad (5)$$

where \cos represents cosine similarity and upsample denotes bilinear interpolation. Note that we omit the softmax for notational simplicity in this expression.

3.3. Architectural Modifications

Self-Self Attention. Recent studies [6, 25, 33, 37, 61] demonstrate that the vanilla implementation has failed to achieve accurate object segmentation. In order to strengthen CLIP's segmentation performance, these studies propose modifying the attention mechanism, resulting in substantial improvement over the vanilla approach. The original self-attention module in CLIP is designed for image-level predictions using x_{CLS} . Because patch tokens do not directly engage with the text features during training, the localization information extracted from attention maps is limited. It is argued that this lack of localization leads to CLIP's suboptimal performance in dense prediction tasks [6, 25]. Thus, we focus on the concept of self-self attention, *e.g.*, key-key (k-k), and query-query (q-q) attentions. Self-self attentions ensure that visual tokens attend to themselves, resulting in higher values both on the diagonal of the attention map and in semantically correlated patches [61]. Compared to the original attention mechanism, self-self attention produces a more spatially localized attention map. During the

experiments, we observed that employing q-q and k-k attentions, similar to SCLIP, yields the best performance. Formally, we use the following attention map formula in the last self-attention block of the image encoder,

$$\begin{aligned} \mathbf{Attn}(X) = & softmax(\frac{XW_QW_Q^TX^T}{\sqrt{d}}) \\ & + softmax(\frac{XW_KW_K^TX^T}{\sqrt{d}}). \end{aligned} \quad (6)$$

Removing Feed-Forward Block. The original Transformer [60] architecture includes FFN layers that process the output of the self-attention block through a series of fully connected layers, enabling the model to capture the complex relationships in the data. ViT used in CLIP’s image encoder also incorporates this feed-forward block. However, due to the inherent structure of pre-training, the parameters of the FFN are optimized for tasks at the image level. Furthermore, CLIPSurgery [37] finds that the FFN in the last encoder block negatively impacts CLIP’s segmentation performance. Therefore, we propose discarding the feed-forward block in the final layer. Our new residual attention block in this layer can be expressed as,

$$X^{(L)} = X^{(L-1)} + \mathbf{SA}(LN(X^{(L-1)})) \quad (7)$$

where L represents the total number of layers, and \mathbf{SA} denotes the self-attention module that utilizes the modified attention map formula presented in Eq. (6).

Incorporating Middle Layer Attention Maps with the Final Layer. Previous methods rely solely on the final layer’s attention map while ignoring valuable feature representations found in intermediate layers. A recent study [22] demonstrates that attention heads in different layers, particularly in the last few layers, are specialized for various image properties (e.g., one head specializes in shapes, while another focuses on colors). Neglecting encoded information in intermediate layers hinders CLIP from achieving its full potential.

In light of these findings, we analyze the attention maps extracted from different layers by applying the modified attention formula described in Eq. (6). Fig. 3 illustrates these attention maps across various layers for a selected patch. We can observe that attention maps from shallow layers tend to attend only to the given patch position. Nevertheless, as we progress from shallow to intermediate layers, attention maps begin to highlight semantically correlated regions related to the selected patch. Although attention maps from the layers just before the final layer may lose certain spatial details, the attention map of the final layer provides the most informative map for the given patch (e.g., the shape of the bear is clearly visible in the final layer). This observation explains why methods that rely solely on the final layer’s attention map can achieve effective results.

On the other hand, this observation also indicates that intermediate layers contain rich information about the image. To leverage this valuable knowledge, we compute two distinct attention maps: the first directly from the final layer and the second by averaging the attention maps from selected intermediate layers. Subsequently, we average these two maps to produce our refined attention map. Formally,

$$\mathbf{Attn}(X) = (\frac{\mathbf{Attn}_L(X) + mean(\mathbf{Attn}_{l'}(X))}{2}) \quad (8)$$

where $\mathbf{Attn}_L(X)$ and $\mathbf{Attn}_{l'}(X)$ represent the attention maps for input X from the final layer L and selected intermediate layers l' , respectively. $mean$ denotes the mean operation across layers, and $\mathbf{Attn}(X)$ is our refined attention map. A detailed analysis of the selected intermediate layers is presented in Sec. 4.3.

3.4. LLM-based Auxiliary Text Generation

With the emergence of large foundational VLMs such as CLIP, models have begun to understand texts beyond the class names in a closed set, allowing for incorporating auxiliary texts alongside original class names. Furthermore, LaCLIP [21] leverages the In-Context Learning (ICL) [7, 46] capability of LLMs to rewrite image captions. It is then trained with these augmented texts, resulting in superior zero-shot classification performance compared to CLIP. Inspired by the success of LaCLIP, we introduce a systematic approach to generate auxiliary texts for open-vocabulary semantic segmentation. We argue that this systematic strategy is more suitable for open-vocabulary settings than manually selecting auxiliary texts since LLMs perform well with classes beyond the predefined set.

To fully leverage the text encoder of CLIP, we generate a synonym and a definition for each given class using LLaMa 3 [19] due to its strong ICL capabilities and open-source nature. For each dataset, we select either synonyms or definitions as the auxiliary text type based on their segmentation performance. Subsequently, we utilize CLIP’s text encoder for original class names and auxiliary texts to obtain text embeddings X^{text} and X_{aux}^{text} , respectively. Lastly, we combine these two text embeddings through a weighted summation as follows,

$$X_{refined}^{text} = \alpha X_{aux}^{text} + (1 - \alpha)X^{text} \quad (9)$$

where $X_{refined}^{text}$ represents the resulting text features, and α is the auxiliary text coefficient.

3.5. Image Engineering

Since CLIP is trained on complete captions, several studies [10, 24, 33, 63, 70] employ prompt templates—an ensemble of different prompts such as "a photo of the {class name}."—to input *prompt-engineered* captions into CLIP’s text encoder. This *Prompt Engineering* strategy

Method	Post-process	COCO-Stuff	COCO-Object	VOC	Context	Cityscapes
Baseline	-	7.1	8.6	20.3	9.0	6.9
ReCo [56]	-	14.8	15.7	25.1	19.9	21.6
GroupViT [65]	-	15.3	27.5	52.3	18.7	18.5
TCL [10]	PAMR	19.6	30.4	55.0	30.4	23.5
MaskCLIP [70]	-	14.6	20.6	38.8	23.2	24.9
CLIP-DIY [63]	-	-	31.0	59.9	-	-
ClearCLIP [33]	-	23.9	33.0	51.8	32.6	30.0
SCLIP [61]	PAMR	23.9	32.1	61.7	31.5	34.1
NACLIP [25]	PAMR	<u>25.7</u>	36.2	64.1	<u>35.0</u>	<u>38.3</u>
TagCLIP* [42]	-	18.7	33.5	64.8	-	-
CaR [58]	Dense-CRF	-	<u>36.6</u>	<u>67.6</u>	30.5	-
ITACLIP (Ours)	PAMR	27.0	37.7	67.9	37.5	40.2

Table 1. **Comparison of ITACLIP with state-of-the-art methods (mIoU, %).** We indicate which post-processing method has been applied to each model, if applicable. * denotes our reimplementation of this model on the COCO-Stuff and COCO-Object datasets. Note that the original paper of TagCLIP [42] evaluates the model on 27 mid-level categories of COCO-Stuff rather than on all 171 classes. Hence, we re-evaluate TagCLIP on COCO-Stuff using all class names for a fair comparison. For each dataset, **bold** values highlight the best scores, while underlined values signify the second-best scores.

enables models to fully leverage the text encoder and improve performance, as demonstrated in [24]. However, previous studies have not applied a similar approach to enhance the input representations for the image encoder, only feeding the original image. Therefore, we propose the *Image Engineering* module, which incorporates data augmentation techniques to expand the input representation. Note that, from this point forward, we omit the [CLS] token to avoid redundancy, as the vanilla approach does not use the [CLS] token for segmentation (see Sec. 3.2). Thus, X^{visual} and its attention map can be represented as $X_{patch}^{visual} \in \mathbb{R}^{N \times d}$ and $\text{Attn}(X) \in \mathbb{R}^{N \times N}$, respectively.

We categorize image augmentations into two groups based on whether they change the spatial structure of the image. Augmentations in the first category maintain the spatial arrangement of the image unchanged, whereas second category augmentations alter the spatial order of the image, requiring a reversal of the augmentation to preserve spatial information. We apply 1) Gaussian blur and 2) grayscale augmentations for the first category, while 1) horizontal and 2) vertical flips are utilized for the second category. Thus, each input image generates four augmentations.

We perform two separate calculations for each augmentation category. Employing our revised image encoder, we compute the first-category visual features X_1^{visual} as,

$$X_1^{visual} = \frac{1}{K+1} \sum_{i=0}^K \frac{X_1^{visual,i}}{\|X_1^{visual,i}\|_F} \quad (10)$$

where $\|\cdot\|_F$ corresponds to the Frobenius norm across the d dimension. $X_1^{visual,i}$ denotes the visual features from the i^{th} first-category augmentation of a total of $K = 2$ augmentations and $X_1^{visual,0}$ is the image itself.

Attention Combination	VOC
q-k	19.0
q-q	58.9
k-k	52.2
v-v	57.7
q-q + k-k	67.9
q-q + v-v	64.9
q-q + k-k + v-v	66.4

Table 2. **Self-self attention combinations.** We evaluate our method with different self-self attention combinations on Pascal VOC. v-v represents the value-value attention.

For the second-category augmentations, we feed augmented images into the image encoder, excluding the original image itself this time. Since the spatial order of the image has been altered, we cannot directly incorporate second-category visual features with X_1^{visual} . Instead, we separately calculate the logits from first and second-category augmentations, namely L_1 and L_2 . $L_1 \in \mathbb{R}^{C \times H \times W}$ is obtained directly by multiplying the embeddings X_1^{visual} and $(X_{refined}^{text})^T$. We then apply a similar operation to all second-category augmented images, generating corresponding logits for each image. To restore the original spatial order, we reverse these augmentations (e.g., a horizontally flipped image is flipped horizontally again). After this reversal, we average the resulting logits to obtain second-category logits. $L_2 \in \mathbb{R}^{C \times H \times W}$. Lastly, we combine L_1 with L_2 through a weighted summation. Formally,

$$L = \lambda L_1 + (1 - \lambda) L_2 \quad (11)$$

where $L \in \mathbb{R}^{C \times H \times W}$ represents the refined image-

FFN	Stuff	Object	VOC	Context	City
✓	26.3	36.9	66.3	36.3	39.4
✗	27.0	37.7	67.9	37.5	40.2

Table 3. **Removing the feed-forward block.** “Stuff”, “Object”, and “City” refer to COCO-Stuff, COCO-Object, and Cityscapes.

engineered logits, and λ is introduced as the image engineering coefficient. Subsequently, we generate the segmentation map S using our refined logits, following the baseline method outlined in Sec. 3.2.

4. Experiments

4.1. Experimental Setup

Datasets. For a fair comparison with previous studies, we evaluate our method on five common semantic segmentation benchmarks using their official validation sets: COCO-Stuff [8], COCO-Object [40], Pascal Context [47], Pascal VOC [20], and Cityscapes [12]. Specifically, COCO-Stuff comprises 171 classes without an explicit background class. COCO-Object is derived from the COCO-Stuff dataset, combining all stuff classes in COCO-Stuff into a single background class and including 80 thing classes from the original dataset. Pascal Context consists of 59 categories and one background class. Pascal VOC contains 20 object classes in addition to one background class. The Cityscapes dataset focuses on urban street scenes, with the validation set containing 19 semantic categories and no separate background class. The validation splits of these datasets include 5000, 5000, 5104, 1449, and 500 images, respectively. We assess the segmentation performance of our method using the mean Intersection over Union (mIoU) metric.

Implementation Details. We use the CLIP-ViT-B/16 [49] model with OpenAI’s pre-trained weights. We define a set of potential background classes for the Pascal VOC and COCO-Object datasets, similar to previous works [42, 58, 61]. This set is fed into CLIP’s text encoder to generate background representations. Additional details about the background set are provided in the Appendix. We utilize ImageNet [13] prompt templates used in CLIP to construct *prompt-engineered* texts. Following previous studies [25, 58, 61], we apply the text expansion technique that uses additional captions for specific classes (*e.g.*, for “person” class: person, person in shirt, person in dress). Furthermore, we perform slide inference with a 224×224 window and a stride of 28, after resizing the short side of images to 336 (560 for Cityscapes), following the procedure described in [61]. Finally, we employ Pixel-Adaptive Mask Refinement (PAMR) [4] to reduce noise in our predictions.

Baselines. We adopt the vanilla approach described in Sec. 3.2 as our baseline method, preserving the original implementation details while omitting post-processing. Also, we compare ITACLIP with other training-free seman-

Intermediate Layers (l')	VOC
✗	65.0
{7}	65.4
{8}	65.5
{7, 8}	65.5
{7, 8, 10}	65.6
{7, 8, 9, 10}	65.5

Table 4. **Impact of selected intermediate layers.** We assess ITA-CLIP with various intermediate layers on Pascal VOC. ✗ indicates that the model does not use intermediate layers for evaluation.

PAMR	Stuff	Object	VOC	Context	City
✗	26.3	36.4	65.6	36.0	39.2
✓	27.0	37.7	67.9	37.5	40.2

Table 5. **Effect of post-processing operation.** Comparing the performance of ITACLIP with and without PAMR on all datasets.

tic segmentation models, including MaskCLIP [70], CLIP-DIY [63], ClearCLIP [33], SCLIP [61], NACLIP [25], CaR [58], and TagCLIP [42], along with weakly-supervised approaches, namely ReCo [56], GroupViT [65] and TCL [10].

4.2. Main Results

Tab. 1 presents the segmentation performance of ITACLIP compared to other approaches. ITACLIP outperforms current state-of-the-art (SoTA) methods on the COCO-Stuff, COCO-Object, Pascal Context, and Cityscapes datasets by a notable margin. In addition, it achieves state-of-the-art performance on Pascal VOC, though with a relatively narrow margin. Furthermore, we emphasize that our method is more robust across all five datasets, whereas most other models experience a performance drop on at least one dataset. For instance, the CLIP-ES-based [41] methods listed in this table—CaR [58] and TagCLIP [42]—fail to maintain their segmentation capability on datasets containing numerous “stuff” classes (*e.g.*, see the results of CaR on Pascal Context and TagCLIP on COCO-Stuff). This performance degradation limits the applicability of these models in real-world scenarios.

In addition to training-free models, we compare our approach with weakly-supervised models, including GroupViT [65] and TCL [10]. We observe that ITACLIP even outperforms these weakly-supervised models without any supervision. We also report the baseline model’s scores in Tab. 1 to illustrate the significant performance gap and demonstrate the need for more advanced approaches.

4.3. Ablation Study

Self-self attention combinations. Self-self attentions yield a more spatially localized attention map, as detailed in

Method	LTG	IE	Context	Stuff
ITACLIP	✗	✗	34.3	24.5
ITACLIP	✓	✗	34.6	24.8
ITACLIP	✓	✓	35.4	25.4

Table 6. **Effect of Image Engineering and LLM-based Text Generation modules.** LTG and IE represent the LLM-based Text Generation and Image Engineering modules, respectively.

Sec. 3.3. To identify the optimal self-self attention combination, we perform an ablation study as presented in Tab. 2. Our method performs best using q-q and k-k attentions with a sum operation. We also provide the score of the original attention formula (q-k attention) to underscore the necessity of self-self attention.

Removing the feed-forward block. In Tab. 3, we investigate the effect of removing the feed-forward block. We observe that FFN in the last layer leads to a drop in segmentation performance across all datasets. The results support the rationale detailed in Sec. 3.3 for removing the FFN.

Study on selected intermediate layers (l'). We analyze the impact of various layers to determine which intermediate layers yield the best performance; Tab. 4 summarizes our findings. To clearly observe the effect of layers, we evaluate the model without PAMR. Based on the visualization in Fig. 3 and a recent study [22] showing that attention maps from the last few layers have a greater effect on the representation, we concentrate more on the middle-deep layers in our analysis. Empirically, employing the 7th, 8th, and 10th layers as intermediate layers results in superior performance. We also report the score using only the final layer’s attention map to demonstrate the effectiveness of our multi-layer approach.

Effect of Image Engineering and LLM-based Text Generation modules. We conduct an ablation study to assess the contributions of the Image Engineering and LLM-based Auxiliary Text Generation modules, as illustrated in Tab. 6. To better understand the impact of these modules, we evaluate our method without applying PAMR. ITACLIP demonstrates superior performance when both modules are employed. This result validates the hypothesis that enriched input features, derived from the Image Engineering and LLM-based Auxiliary Text Generation modules, can enhance image segmentation performance.

Effect of post-processing operation. Before generating the final predictions, we apply PAMR to mitigate noise in our results. We examine the effect of PAMR on our method across all datasets, as shown in Tab. 5. Although PAMR improves segmentation performance across all datasets, ITACLIP still achieves competitive results and even surpasses the SoTA on some datasets without PAMR. These results demonstrate that our method is highly robust and performs well without relying heavily on post-processing.

Stride	Stuff	Object	VOC	Context	City
224	25.3	36.7	66.1	36.9	37.7
112	26.6	37.4	67.1	37.4	39.7
56	26.9	37.7	67.9	37.5	40.2
28	27.0	37.7	67.9	37.5	40.2

Table 7. **Role of stride value.** We investigate the role of the stride value in our method across all five datasets.

λ	α	Context
0.75	0.15	36.0
0.6	0.15	35.9
0.5	0.15	35.8
0.75	0.2	35.9
0.6	0.2	35.9
0.6	0.1	35.9
0.5	0.1	35.8

Table 8. **Influence of Hyperparameters.** λ and α denote the Image Engineering and Auxiliary Text Coefficients, respectively. The experiments are conducted without applying PAMR.

Role of stride value. In Tab. 7, we investigate the impact of different stride values on our method. Our results indicate that lower stride values tend to improve segmentation performance. Considering that lower stride values increase computational cost, we also report ITACLIP’s segmentation performance using higher stride values to enable faster inference. We observe that ITACLIP’s performance with a stride of 56 nearly matches the results obtained using our default stride of 28. Even with a stride of 112, our method achieves state-of-the-art results on most datasets.

Influence of Hyperparameters. Tab. 8 illustrates the effect of Image Engineering (λ) and Auxiliary Text Coefficients (α) on the Pascal Context dataset. We observe minimal performance variance within the range of values listed in the table, highlighting the robustness of our method.

5. Conclusion

This study introduces ITACLIP, which leverages CLIP’s image-level knowledge for dense prediction tasks without requiring pixel-level annotations or additional training. We propose architectural modifications to CLIP’s visual encoder and introduce the Image Engineering and Auxiliary Text Generation modules to expand the input representation. As a result of these modifications and modules, our experiments show that ITACLIP achieves SoTA results across five segmentation datasets. Additionally, the Image Engineering module and LLM-based Text Generation strategy can be seamlessly integrated into a range of computer vision tasks. These modules will assist researchers in further improving prediction quality across various vision tasks.

References

- [1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *ICCV*, pages 1348–1357, 2023. 1
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 2
- [3] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 33:25–37, 2020. 2
- [4] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pages 4253–4262, 2020. 7
- [5] Luca Barsellotti, Roberto Amoroso, Lorenzo Baraldi, and Rita Cucchiara. Fossil: Free open-vocabulary semantic segmentation through synthetic references retrieval. In *WACV*, pages 1464–1473, 2024. 1, 2
- [6] Walid Bousseth, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, pages 3828–3837, 2024. 2, 4
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 5
- [8] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 1, 2, 7, 12
- [9] Cambridge University Press. Cambridge dictionary. <https://dictionary.cambridge.org/>. Accessed: 2024-08-24. 12
- [10] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, pages 11165–11174, 2023. 3, 5, 6, 7
- [11] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 2
- [12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 7, 12
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7
- [14] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 2
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2
- [16] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022. 2
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3
- [18] Yu Du, Fangyun Wei, Zihe Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pages 14084–14093, 2022. 1
- [19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5, 12
- [20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 2, 7, 12, 13
- [21] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 5
- [22] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip’s image representation via text-based decomposition. In *ICLR*, 2024. 5, 8
- [23] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022. 3
- [24] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 5, 6
- [25] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2404.08181*, 2024. 1, 2, 4, 6, 7, 12, 13
- [26] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *ICCV*, pages 1086–1096, 2023. 3
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3
- [28] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets. *arXiv preprint arXiv:2304.09324*, 2023. 2
- [29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Han-

- nanah Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 1, 2
- [30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 2
- [31] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *NeurIPS*, 36:35631–35653, 2023. 3
- [32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2
- [33] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv preprint arXiv:2407.12442*, 2024. 1, 2, 4, 5, 6, 7, 12
- [34] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3
- [35] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021. 2
- [36] Yunheng Li, Zhongyu Li, Quansheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *ICML*, 2024. 1
- [37] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 1, 2, 4, 5
- [38] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, pages 7667–7676, 2023. 3
- [39] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 3
- [40] 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 *ECCV*. Springer, 2014. 2, 7, 12, 13
- [41] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, pages 15305–15314, 2023. 1, 3, 7
- [42] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *AAAI*, volume 38, pages 3513–3521, 2024. 1, 3, 6, 7, 12
- [43] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 33:11525–11538, 2020. 3
- [44] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, pages 23033–23044. PMLR, 2023. 3
- [45] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 2
- [46] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *NAACL-HLT*, 2022. 5
- [47] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 2, 7, 12, 13
- [48] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2018. 2
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4, 7
- [50] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *Technical report, OpenAI*, 2018. 2
- [51] 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. 2
- [52] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *ICCV*, pages 5571–5584, 2023. 1, 3
- [53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 2
- [54] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *ECCV*. Springer, 2024. 1
- [55] Cheng Shi and Sibe Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *ICCV*, pages 15724–15734, 2023. 1
- [56] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *NeurIPS*, 35:33754–33767, 2022. 1, 6, 7
- [57] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping

- Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *ICCV*, pages 15453–15465, 2023. [1](#)
- [58] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, pages 13171–13182, 2024. [2](#), [3](#), [6](#), [7](#), [12](#)
- [59] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *CVPR*, pages 14214–14223, 2023. [1](#)
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. [5](#)
- [61] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. [1](#), [2](#), [4](#), [6](#), [7](#), [12](#), [13](#)
- [62] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022. [1](#)
- [63] Monika Wysoczańska, Michaël Ramamonjisoa, Tomasz Trzciński, and Oriane Siméoni. Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free. In *WACV*, pages 1403–1413, 2024. [2](#), [3](#), [5](#), [6](#), [7](#)
- [64] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. [2](#)
- [65] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022. [3](#), [6](#), [7](#)
- [66] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *CVPR*, pages 2935–2944, 2023. [3](#)
- [67] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, pages 2955–2966, 2023. [3](#)
- [68] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*. Springer, 2022. [2](#)
- [69] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. [3](#)
- [70] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer, 2022. [2](#), [3](#), [5](#), [6](#), [7](#)
- [71] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *CVPR*, pages 11175–11185, 2023. [1](#), [2](#)