Image-to-Image Matching via Foundation Models: A New Perspective for Open-Vocabulary Semantic Segmentation

Yuan Wang[1*] Rui Sun[1*] Naisong Luo[1] Yuwen Pan[1] Tianzhu Zhang[1,2†]
[1]University of Science and Technology of China
[2]Deep Space Exploration Laboratory
{wy2016, issunrui, lns6, panyw}@mail.ustc.edu.cn, {tzzhang}@ustc.edu.cn

Abstract

Open-vocabulary semantic segmentation (OVS) aims to segment images of arbitrary categories specified by class labels or captions. However, most previous best-performing methods, whether pixel grouping methods or region recognition methods, suffer from false matches between image features and category labels. We attribute this to the natural gap between the textual features and visual features. In this work, we rethink how to mitigate false matches from the perspective of image-to-image matching and propose a novel relation-aware intra-modal matching (RIM) framework for OVS based on visual foundation models. RIM achieves robust region classification by firstly constructing diverse image-modal reference features and then matching them with region features based on relation-aware ranking distribution. The proposed RIM enjoys several merits. First, the intra-modal reference features are better aligned, circumventing potential ambiguities that may arise in cross-modal matching. Second, the ranking-based matching process harnesses the structure information implicit in the inter-class relationships, making it more robust than comparing individually. Extensive experiments on three benchmarks demonstrate that RIM outperforms previous state-of-the-art methods by large margins, obtaining a lead of more than 10% in mIoU on PASCAL VOC benchmark.

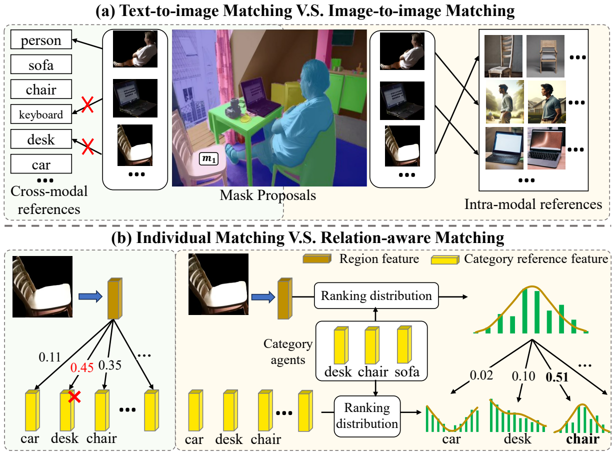

Figure 1. Motivation of our method. (a) False matches tend to occur in cross-modal features. We establish well-aligned image-modal reference features thus transit the text-to-image matching to image-to-image matching. (b) Indivdual matching tends to suffer from disturbances. We propose a novel relation-aware matching strategy for more robust region classification.

1. Introduction

Aiming at allocating semantic labels to the corresponding pixels, semantic segmentation has achieved conspicuous achievements attributed to the development of large-scale datasets [11, 22, 25, 74] and elaborate algorithms [7, 8, 28]. However, the capabilities of conventional semantic segmentation models are restricted to predefined training categories, failing to recognize a broader spectrum of concepts, which poses a severe limitation for their practical applications. In pursuit of the human-like intelligence of unbounded and fine-grained scene understanding, open-vocabulary segmentation (OVS) [2, 67] has attracted increasing interest recently, which enables segmentation with an unrestricted vocabulary.

OVS is highly challenging as it requires not only grouping all pixels to corresponding visual concepts but also assigning each region the correct semantic label from a large vocabulary. Some early attempts [29, 64] optimize the pixel grouping process and the classification of visual concepts simultaneously by employing contrastive learning with image-text pairs. These methods inevitably yield masks of relatively low quality with only coarse supervision available. Another line of works [24, 67, 68] significantly dominate this field by modeling the OVS as a region recognition problem, which decouple OVS into two procedures, *i.e.* class-agnostic mask proposals prediction and mask class recognition. Though the mask proposals generation has been significantly improved with the assistance of pixel-level supervision and

[*]Equal contribution
[†]Corresponding author

elaborate segmenter architectures (*e.g*. Mask2Former [8]), the unsatisfactory region class recognition has emerged as the performance bottleneck of OVS [24]. We attribute this to the natural gap between the highly abstract and monotonous category textual features and the visual features that are more concrete and diverse.

Most current OVS approaches achieve mask class recognition via cross-modal region-category alignment based on the pretrained vision-language models (*e.g*., CLIP [38]). Despite its exemplary generalization ability on downstream classification tasks, CLIP usually suffer from spatial relations ambiguity [66, 69] and co-occurring object confusion ascribed to the holistic pre-training objective. On the other side, also pre-trained on internet-scale data, Stable Diffusion [42] (SD) model has garnered growing research interest credited to the phenomenal power of synthesizing photorealistic images with diverse and plausible content conditioned on textual descriptions. Considering that features from the same modality inherently exhibit a higher degree of alignment compared to cross-modal features (as shown in Fig. 6), we ask the question: ❶ *is it possible to transition the region classification in OVS from image-to-text matching to image-to-image matching using the SD model?*

Before tackling this issue, we first revisit the common practices in mask class recognition [24, 67], which involves matching visual concepts with a set of category reference features (*e.g*. category textual features from CLIP). Inspired by previous exploration [59, 60, 73] of the strong image-text correspondence exhibited by the SD model, we can establish better-aligned intra-modal category reference features as shown in Fig. 1 (a). The cross-attention maps in the conditional generation process could serve as exceptional tools to further refine the category features. Nonetheless, it is still non-trivial to map test region features to their corresponding category reference features precisely, as the intra-class diversity and disturbances from akin categories exhibit in practical scenarios potentially lead to mismatches. Previous methodologies process each category independently during the matching phase, overlooking the informative inter-class relationships, which implicitly integrate structured contextual information and effectively aid in disambiguation. Here another question arise: ❷ *how to harness the structure information modeled in the inter-class relationship to facilitate more accurate matching?*

Driven by this question, we carefully design the relation-aware similarity measurement that incorporates the relations of the current region with some semantic relevant category agents. For a specified region feature within test image, a subset of category reference features that most akin to the region are selected to serve as category agents. The central idea is that we regard the category agent ranking as a stochastic event rather than a deterministic permutation. For example, in ranking, given the region feature specified by the mask proposal m_1 in Fig. 1 (a), its scores vary for different category agents, which can be taken as probabilities. The probability of being ranked first is 0.45 of the agent 'desk' and 0.35 of the agent 'chair'. The ranking permutation reflects the relevance of the corresponding categories w.r.t. the region feature. An agent-ranking probability distribution can be constructed by associating the probability with every rank permutation for both the region feature and all category reference features. Finally, as illustrated in Fig. 1 (b), we transform the similarity measurement from individual region-reference comparison to relation-aware agent-ranking distribution similarity.

In this paper, we present a training-free OVS framework RIM to achieve **R**elation-aware **I**ntra-modal **M**atching based on visual foundation models that coherently addresses the questions ❶-❷. Specifically, we facilitate a synergistic collaboration between the SD model and SAM to generate category-specific reference features by prompting the Segment Anything Model [22] (SAM) with points selected from the cross-attention map within SD model. SAM is further adopted to provide mask proposals of the test images. Finally, we conduct the relation-aware matching based on ranking distribution in the robust all-purpose feature space of DINOv2 [35]. RIM ensembles expertise of different visual foundational models in a complementary manner, enhancing mask quality and region-category matching precision simultaneously. Moreover, the overall framework is training-free, substantially mitigating the risk of overfitting.

Our contributions can be concluded as follows: (1) We reveal the problems of region feature classification in OVS and propose to construct well-aligned intra-modal reference features to circumvent ambiguities of cross-modal matching. (2) We design a relation-aware matching strategy based on ranking distribution, which captures structure information implicit in inter-class relationships and enables more robust matching. (3) We propose a training-free relation-aware intra-modal matching (RIM) network for OVS based on visual foundation models. Extensive experiments demonstrate that RIM remarkably surpasses previous state-of-the-art methods by a large margin.

2. Related Work

Semantic Segmentation is a fundamental computer vision task with widespread applications in fields such as medical image processing [31, 37, 46, 48, 49, 58], video analysis [47, 51]. The pioneering FCN [28] has inspired a multitude of subsequent endeavors [6, 7, 43, 62, 72] centered around Convolutional Neural Networks (CNN). Beyond the CNN-based models, the triumph of the Vision Transformer (ViT) has spurred a succession of transformer-based segmentation models [30, 45, 63, 71], which have progressively evolved into a unified segmentation framework [8, 9] capable of addressing various segmentation tasks. Some alternative settings, such as few-shot [23, 55–57], semi-supervised [23, 32, 33, 50], or weakly supervised [19, 54] segmentation, attempt to enhance the practicality of semantic segmentation techniques. Despite their success, these models are confined to predefined training categories or the specific foreground class, failing to recognize a broader spectrum of categories.

Open-vocabulary Semantic Segmentation aims to segment images with arbitrary categories described by texts. Some efforts [2, 29, 61, 65, 69] concentrated on developing a joint embedding space that bridges image pixels with class names or descriptions via learning objectives established from image-text pairs. For example, GroupViT [64] designed a hierarchical grouping transformer and learned the alignment between groups and text via contrastive loss. Another line of works [24, 66, 67] models OVS as a region recognition task by decoupling it into mask proposal generation and region classification, achieving notable progress attributed to advanced segmentation architectures [8] and pixel-level annotations [25, 74]. Large-scale vision-language pre-training models such as CLIP [38] and ALIGN [20] have been widely applied for OVS [15, 68, 70] which endows OVS model enhanced generalization. However, those pre-training models often encounter spatial confusion in dense prediction tasks [69], leading to recurring misclassification issues, which has emerged as a critical bottleneck in OVS.

Visual foundation models is catching up with the research in natural language processing (NLP) [1, 10, 36] and have achieved conspicuous achievements across a wide range of visual tasks. For example, DINOv2 [35] establishes an impressive all-purpose feature extractor via self-supervised learning at both the image and patch levels. Trained on over 1 billion masks, Segment Anything Model (SAM) [22] has demonstrated astonishing zero-shot class-agnostic segmentation performance. Diffusion models [18, 44] have propelled the advancement of a series of image-to-text generation systems such as DALL-E [39] and Stable Diffusion model [42]. A series of works [27, 60, 66] render visual foundation models as powerful out-of-the-box tools to handle downstream tasks. Though a single foundation model may have limited capacity in addressing complex visual tasks such as OVS, in this work, we demonstrate that integrating different foundation models leads to positive synergies.

3. Method

3.1. Problem Definition

Open-vocabulary segmentation aims to segment any image against a new vocabulary of categories $\mathbf{C}_{test}$. Previous approaches typically involve training models on datasets annotated with image-level textual labels or pixel-level masks, with the category set $\mathbf{C}_{train}$. $\mathbf{C}_{test}$ contains novel categories not exposed to the training process, *i.e.*, $\mathbf{C}_{train} \neq \mathbf{C}_{test}$. Owing to the training-free nature of our method, it inherently operates under a more challenging zero-shot setting wherein all test categories are considered novel.

3.2. Preliminary of Vision Foundation Models

Stable-Diffusion model. Different from discriminative image-text models that model the class probability distribution $\boldsymbol{p}(\boldsymbol{c}|\boldsymbol{i})$ given the image, text-to-image stable diffusion model encode a text-conditional distribution of possible images $\boldsymbol{p}(\boldsymbol{i}|\boldsymbol{c})$. This model demonstrates the prowess to synthesize high-fidelity images $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ that closely align with the specified conditional text $\mathcal{T}$, all originating from a latent space defined by random Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, 1)$. Specifically, stable diffusion model consists of three components: a pre-trained variational autoencoder (VAE) [13] that encodes and decodes the image latent code, a text encoder for prompt embedding, and a time-conditional UNet [43] for the denoising of latent vectors. The visual-text interaction occurs in the cross-attention layers that are integral to the UNet structure for each denoising step. Target categories within synthesized images can be accurately localized based on class-discriminative cross-attention maps.

Segment Anything Model. SAM comprises of three components: an image encoder $\mathbf{Enc_I}$, a prompt encoder $\mathbf{Enc_P}$, and fast mask decoder $\mathbf{Dec_M}$. $\mathbf{Enc_P}$ takes as input the prompts of various optional forms such as points, boxes or coarse masks, and translates the prompts into feature tokens $\mathbf{F_P}$. Image feature tokens $\mathbf{F_I}$ are also obtained from $\mathbf{Enc_I}$. A series of learnable mask tokens $\mathbf{F_M}$ are introduced and concatenated with $\mathbf{F_P}$ for diverse masks generation. Then the light weight $\mathbf{Dec_M}$ integrates the $\mathbf{F_I}$ and the concatenation of $\mathbf{F_P}$ and $\mathbf{F_M}$ to predict segmentation masks.

DINOv2. Pretrained on a large quantity of curated data with image and patch level distriminative self-supervised learning, DINOv2 [35] learns all-purpose visual features that work out of box on many tasks varing from image level, *e.g.*, image classification and pixel level, *e.g.*, semantic segmentation. Moreover, this all-purpose ViT model demonstrates impressive patch-matching ability, robustly capturing similar semantic intent across different objects and even distinct images [27].

3.3. Overview

As shown in Fig. 2, we propose a training-free OVS framework RIM based on visual foundation models. RIM tackles the challenging region classification in OVS from a novel image-to-image matching perspective via the following two procedures, *i.e.*, 1) intra-modal reference features construction, and 2) relation-aware matching. We resort to the SD model [42] and the SAM [22] to establish category reference features in procedure 1). In procedure 2), we conduct ranking based matching between the reference features and the region features in the all-purpose DINOv2 feature space. The details are as follows.

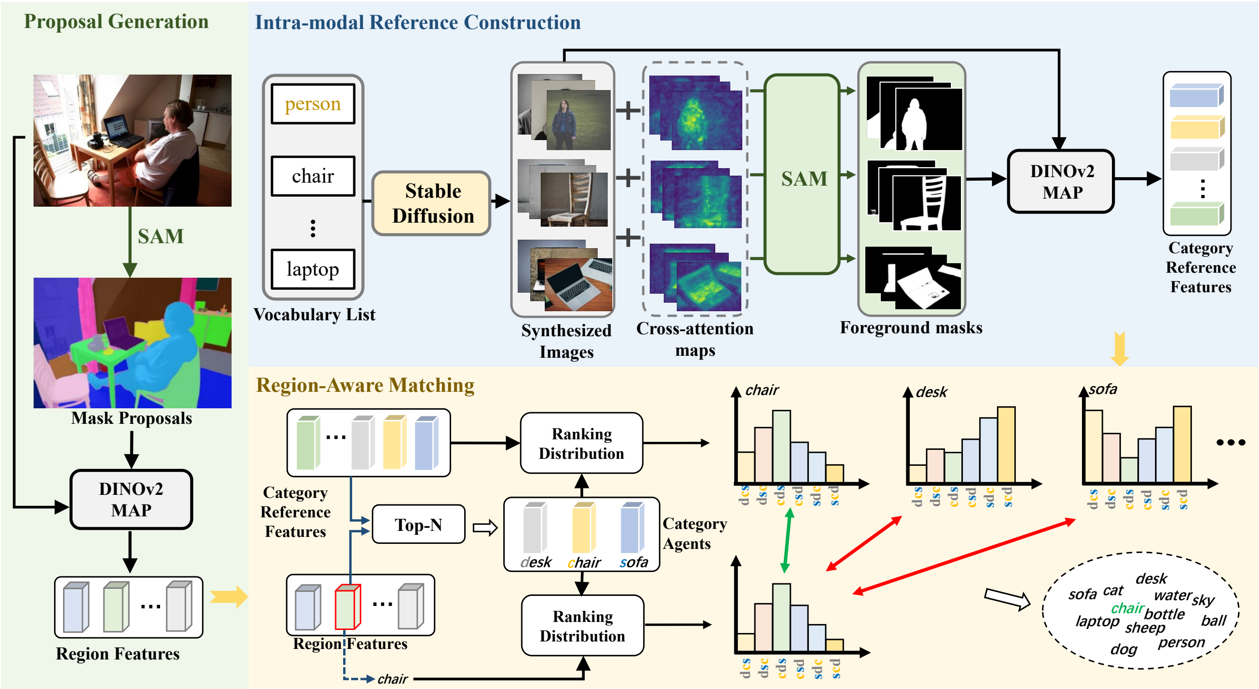

Figure 2. Framework of our proposed Relation-aware Intra-modal Matching (RIM) Network. We first explore Stable Diffusion model and SAM to construct image-modal reference features, then we conduct relation-aware matching between region features and reference features based on ranking distribution. The matching is established in the all-purpose feature spaces of DINOv2.

3.4. Intra-modal Reference Construction

In contrast to the substantial gap between abstract, monotonous text features and the concrete, diverse image features, homogenous modality features naturally exhibit improved alignment characteristics as shown in Fig. 6. We resort to stable diffusion model to generate a handful of reference images $\mathcal{I} = \{\mathbf{I}_1^c, \mathbf{I}_2^c, ..., \mathbf{I}_K^c | \mathbf{I}_k^c \in \mathbb{R}^{H\times W\times 3}, c = 1, 2, \dots, C\}$ for all C candidate classes by simply prompting it with "a photo of [category name]". Together with the images, we also acquire the corresponding cross-attention maps for localizing the foreground targets. Specifically, for time step t, the noisy image features $\mathbf{F}_v \in \mathbb{R}^{h\times w\times c}$ are flattened and linearly projected into the $queries$ $\mathbf{Q}$ and the prompt features $\mathbf{F}_p \in \mathbb{R}^{N\times d}$ are respectively projected into the $keys$ $\mathbf{K}$ and $values$ $\mathbf{V}$:

$$\mathbf{Q} = \mathbf{F}_v\mathbf{W}^{\mathcal{Q}}, \quad \mathbf{K} = \mathbf{F}_p\mathbf{W}^{\mathcal{K}}, \quad \mathbf{V} = \mathbf{F}_p\mathbf{W}^{\mathcal{V}}. \tag{1}$$

Among which, $\mathbf{W}^{\mathcal{Q}}$, $\mathbf{W}^{\mathcal{K}}$, $\mathbf{W}^{\mathcal{V}}$ are linear projections. The cross-attention maps are calculated as:

$$\mathcal{S} = \mathbf{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d}}), \tag{2}$$

where the $\sqrt{d}$ is the scaling factor. Cross-attention maps $\mathcal{S}_n^{l,t}$ of different text tokens from different layers of UNet can be obtained based on the Equ. 2 and n,l,t are the index of text tokens, UNet layers and diffusion steps, respectively. To obtain the cross-attention map corresponding to the category token for foreground mining, we follow [60] to aggregate normalized multi-time and multi-layer maps, formaly,

$$\bar{\mathcal{S}}_n = \frac{1}{L\cdot T}\sum_{l\in L, t\in T}\frac{\mathcal{S}_n^{l,t}}{\mathbf{max}(\mathcal{S}_n^{l,t})}, \tag{3}$$

Where the L and T denote the total time steps and the number of UNet layers.

The foreground area in the synthesized images should be further located to avoid the interference of the irrelevant background regions. However, simply binarizing the cross-attention maps by thresholding may fail to fully capture the target. We thus exploit SAM to generate target masks by sampling the prompt points within the binarized attention maps. The foreground masks $\mathcal{M} = \{\mathbf{M}_1^c, \mathbf{M}_2^c, ..., \mathbf{M}_K^c | \mathbf{M}_k^c \in \mathbb{R}^{h\times w}, c = 1, 2, \dots, C\}$ and corresponding images $\mathcal{I}$ then serve as the class references.

To enhance the robustness of instance-level matching, we opt to construct class reference image features within the all-purpose feature space of DINOv2 [35] via mask average pooling (MAP). Specifically,

$$\mathcal{F}_{ref} = \{\mathbf{F}^c | \mathbf{F}^c = \frac{1}{K}\sum_{k\in K}\mathbf{MAP}(\varphi(\mathbf{I}_k^c), \zeta(\mathbf{M}_k^c))\}_{c=1}^{C} \tag{4}$$

where the φ and ζ denote the DINOv2 feature extractor (ViT) and the bilinear-interpolation resize function, respectively.

The $\mathbf{F}^c \in \mathbb{R}^{1\times D}$ represents the average of all synthesized foreground features. We also construct a background reference feature by averaging all the background features of synthesized images. Similarly, region features within the test image specified by mask proposals are obtained as:

$$\mathcal{F}_{test} = \{\mathbf{F}^p | \mathbf{F}^p = \mathbf{MAP}(\varphi[\mathbf{I}_{test}], \zeta[\mathbf{M}^p])\}_{p=1}^{P}, \tag{5}$$

among which $\mathbf{P}$ is the number of masks generated by SAM. By designing collaborative interactions among various visual foundational models, we establish well refined image-modal reference features thus shifts the paradigm of region classification in OVS from cross-modal matching to better-aligned intra-modal matching.

3.5. Relation-aware Matching

The naive region classification strategy only selectively recruits the class reference feature with the highest similarity. However, independently taking each category exacerbates the risk of mismatches owing to the intra-class diversity and interference from similar categories. To further harness the structure information implicit in inter-class relationship for more effective matching, we select the top N reference features most similar to the region feature $\mathbf{F}^p$ as category agents $\mathbf{A}^p \in \mathbb{R}^{N\times D}$. We derive the scores of region-agent relation $s^p \in \mathbb{R}^{1\times N}$ between the region feature $\mathbf{F}^p$ and the category agents $\mathbf{A}^p$ using cosine similarity:

$$\mathbf{s}^p = \frac{\mathbf{F}^p(\mathbf{A}^p)^\intercal}{\|\mathbf{F}^p\|_2 \cdot \|\mathbf{A}^p\|_2}. \tag{6}$$

The core idea is that we take the agent ranking as a random event rather than a deterministic permutation. This implies that each permutation of the category agents is present with a certain probability, as opposed to the exclusive existence of the permutation ordered from largest to smallest. The probability of one permutation $\pi \in \mathcal{P}(|\mathcal{P}| = N!)$ given s (we omit region index p for brevity) can be calculated as:

$$P(\pi|\mathbf{s}) = \prod_{k=1}^{K} \frac{\mathbf{s}_{\pi(k)}}{\sum_{k'=k}^{K} \mathbf{s}_{\pi(k')}}, \tag{7}$$

among which the $\pi(k)$ denotes the k^{th} class index of this permutation. For instance, assume that for a given region, the selected agents correspond to the categories "sofa", "chair", and "bench" respectively. One of the permutations of these three agents is $\pi = (chair, sofa, bench)$. We can derive the probability of π based on the region-agent relation s:

$$P(\pi|\mathbf{s}) = \frac{\mathrm{s}(\mathit{chair})}{\mathrm{s}(\mathit{sofa}) + \mathrm{s}(\mathit{bench}) + \mathrm{s}(\mathit{chair})} \cdot \frac{\mathrm{s}(\mathit{sofa})}{\mathrm{s}(\mathit{sofa}) + \mathrm{s}(\mathit{bench})}. \tag{8}$$

By associating the probabilities of all $|\mathcal{P}|$ permutations, We convert the scores of individual region-agent relation s into class ranking probability distributions $P(\pi \in \mathcal{P}|\mathbf{s}^p) \in \mathbb{R}^{1\times|\mathcal{P}|}$, effectively modeling the inter-class relationship. Similarly, we compute the agent-ranking probability distributions for each category reference features on the same category agents, resulting in $P(\pi \in \mathcal{P}|\mathbf{s}^r) \in \mathbb{R}^{1\times|\mathcal{P}|}$, where $\mathbf{s}^r$ is obtained through Equ. 6. The distribution of the region is compared against the distributions of all reference features via cosine similarity, enabling the determination of the classification result:

$$\mathit{cls} = \mathop{\arg\max}_{r=1,2,\ldots,C} [\mathbf{cosine}(P(\pi \in \mathcal{P}|\mathbf{s}^p), P(\pi \in \mathcal{P}|\mathbf{s}^r))]. \tag{9}$$

To further leverage the advantage of diversity in image-modal reference features, we construct a series of subcategory reference features by clustering the foreground prototypes (obtained by mask average pooling) of all synthesized images. The subcategory reference features are involved in the similarity computation instead of the holistic ones, and the final score of a category is obtained by summing up the similarities of all the corresponding subcategory reference features. The classification result is then obtained by:

$$\mathit{cls} = \mathop{\arg\max}_{r=1,2,\ldots,C} [\sum_{t=1}^{T} \mathbf{cosine}(P(\pi \in \mathcal{P}|\mathbf{s}^p), P(\pi \in \mathcal{P}|\mathbf{s}_t^r))]. \tag{10}$$

Based on the metric grounded in distributional similarity, instances of erroneous region classification can be effectively reduced as inter-class relationships are exploited to facilitate disambiguation.

4. Experiments

4.1. Dataset and Evaluation Metric

We evaluate our model on three commonly used benchmarks, namely, PASCAL VOC 2012 [14], PASCAL Context [34] and COCO Object [25], which have 20,59,80 foreground classes, respectively. We also consider an extra background class in all three datasets. Training sets of datasets are not needed as the proposed method is training-free. For a fair comparison with previous approaches, we directly evaluate our method on the validation sets of these datasets, including 1449, 5105, and 5000 images, respectively. We use the mean of class-wise intersection over union (mIoU) following the common practice to measure the performance.

4.2. Implementation of Visual Foundation Models

We adopt the Stable Diffusion model v1.4 [42] to generate category-specific images of resolution 512×512 for category reference feature construction. We employ DINOv2 [35] with a ViT-B [12] as the default image encoder for more discriminative matching. Specifically, the $keys$ sequence of the last attention layer is reshaped as the feature map. SAM [22] with ViT-B is adopted as the segmenter. We collect 32x32

Method	Arch	Training dataset	Supervision	Zero-shot transfer	Downstream datasets		
					PASCAL VOC	PASCAL Context	COCO Object
DeiT [52]	ViT	IN-1K	class label	✗	53.0	35.9	-
MoCo [16]	ViT	IN-1K	self	✗	34.3	21.3	-
DINO [3]	ViT	IN-1K	self	✗	39.1	20.4	-
ViL-Seg [26]	ViT	CC12M	self+text	✓	33.6	15.9	-
MaskCLIP [75]	ViT	LAION	$text + CLIP_T$	✓	38.8	23.6	20.6
GroupViT [64]	ViT	CC12M	text	✓	52.3	22.4	-
ZeroSeg [5]	ViT	IN-1K	$CLIP_V$	✓	40.8	20.4	20.2
TCL [4]	ViT	CC3M+CC12M	text	✓	51.2	24.3	30.4
ViewCo [41]	ViT	CC12M+YFCC	text+self	✓	52.4	23.0	23.5
CLIPpy [40]	ViT	HQITP-134M	text	✓	52.2	-	32.0
SegCLIP [29]	ViT	CC3M+COCO	$text + CLIP_T$	✓	52.6	24.7	26.5
OVSegmentor [65]	ViT	CC4M	self+text	✓	53.8	20.4	25.1
SimSeg [69]	ViT	CC3M+CC12M	text	✓	57.4	26.2	29.7
DiffSeg [53]	UNet+ViT	Training-free		✓	60.1	27.5	37.9
OVDiff [21]	UNet			✓	67.1	30.1	34.8
Ours	UNet+ViT			✓	**77.8**	**34.3**	**44.5**

Table 1. **Comparison with existing methods.** Models in the first three rows are finetuned on target datasets while the rest approaches perform zero-shot transfer. Bold fonts refer to the best results among the models which enable zero-shot transfer. With only image-text pairs available, our method significantly outperforms the existing approaches. More results please refer to **Supplementary Material**.

IRC		RM		mIOU
w/o fg-mask	w/ fg-mask	w/o sub	w/ sub	
				26.7
✓				38.7
	✓			41.8
	✓	✓		43.1
	✓		✓	**44.5**

Table 2. Ablation studies of the proposed RIM. We mainly verified the effectiveness of the image reference construction (IRC), and relation-aware matching (RM). Moreover, we also observe the effectiveness of foreground segmentation (fg-mask) of synthesized images and the subcategory reference features (sub).

Encoder	CLIP [38]	SD [42]	MAE [17]	DINOv2
mIoU	42.0	43.1	39.5	**44.5**

Table 3. Comprison of different image feature extractors.

Segmenter	none	MaskFormer	SAM
mIoU	42.0	76.7	**77.8**

Table 4. Comparison of mask proposal generators. "none" means direct pixel classification.

Selection of Agents	mIoU
Rand	43.8
Top-N	**44.5**

Table 5. Comparison of category agents selection.

prompt points in a grid manner to generate mask proposals for the test images. To segment the foreground of synthesized reference images, We sample 5 prompt points within the binarized cross-attention map, where the binarization threshold is set to a relatively high value of 0.7 to prevent erroneous segmentation of background regions. The number of category agents is set to 4. More implementation details are provided in **Supplementary materials**.

4.3. Comparison to the state-of-the-arts

We compare our method with approaches that have been trained with fully supervised finetuning transfer and zero-shot transfer. Besides, we also conduct performance comparisons with newly arising train-free approaches. Table 1 summarizes the results of the comparison.

Firstly, we can observe that the proposed RIM significantly surpasses the non-zero-shot fully supervised segmentation baselines, *i.e.*, DeiT [52], MoCo [16]. Furthermore, we compare with other zero-shot OVS approaches such as MaskCLIP [75], SegCLIP [29], and SimSeg [69], which also adopt the ViT backbones of visual foundation model, *i.e.*, CLIP [38]. Our method also shows a clear lead. Specifically, we achieve 20.4%, 18.1%, 14.8% mIoU over the SimSeg [69] on three datasets, respectively. We posit that the primary cause is that CLIP is susceptive to ambiguities in spatial relations and confusion of co-occurring objects, a consequence stemming from its holistic training objective. Besides, RIM also demonstrates significant performance improvements over the most recently developed training-free methods based on visual foundation models, achieving 10.4%, 4.2%, and 6.6% mIoU gains, respectively. As shown in Fig. 3, we present a series of visualizations depicting segmentation outcomes across various datasets for a more intuitive observation.

4.4. Ablation Study

A series of ablation studies are conducted to thoroughly investigate the impact of each component of the proposed RIM. As shown in Table 2, we mainly implement our experiments on COCO Object [25] dataset, and the first row is our baseline, which follows a naive image-to-text matching for region classification. Specifically, we crop the test image along the bounding boxes corresponding to the mask proposals (gen-

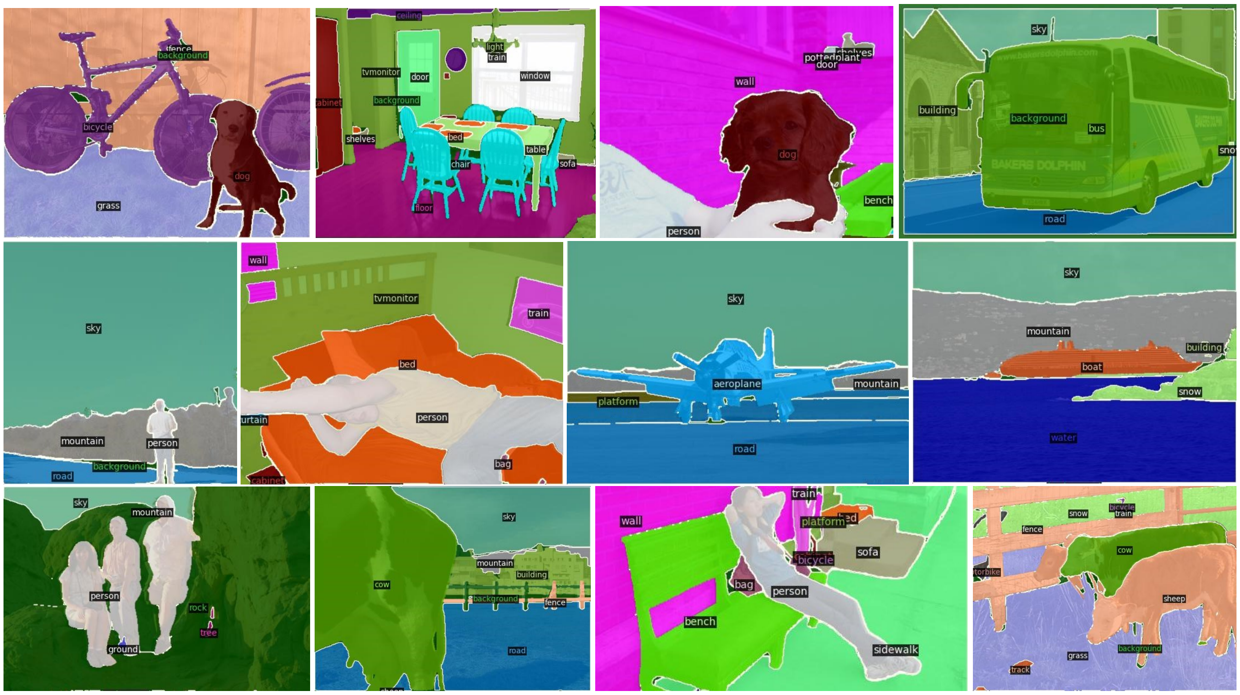

Figure 3. Qualitative results of our method.

erated by SAM) and resize them to 224 × 224 resolution. The background area of the cropped image is filled with zeros, and then CLIP [38] with ViT-B is exploited to perform image-to-text matching between the cropped image and category labels.

Ablation study on intra-modal reference construction. To verify the effectiveness of intra-modal alignment, we first construct a naive image-to-image matching as the 2^{nd} row of Table 2. More concretely, DINOv2 features of all synthesized images of a category are condensed to a holistic reference feature via global average pooling. The matching process is simply implemented with cosine similarity. Despite its simplicity, this intra-modal matching exhibits a significant performance improvement, *i.e.*, 12.0% in mIoU over the cross-modal matching based on CLIP. This improvement is anticipated as the holistic training objective of CLIP makes its prediction heavily reliant on contextual information. This reliance not only introduces spatial confusion but also leads to misclassification of frequently co-occurring objects, such as the sky and $airplane$. Attributed to the image generation capabilities of SD model and the robust high-level semantic feature extraction of DINOv2, an improved intra-modal alignment between region features and class reference features is achieved.

Further performance improvements can be observed if we construct the category reference features with only foreground features as shown in the 3^{th} row of Table 2. We deem the main reason is that the background areas in synthetic images may contain information of other categories, which confuses the classification process. While in our implementation, more refined category reference features are constructed attributed to the spatial localization ability of the SD model embedded in the cross-attention map and the powerful segmentation capability of SAM, which effectively mitigates the influence of cluttered backgrounds. Despite utilizing merely cosine similarity for intra-modal matching, it has already achieved impressive performance, *i.e.*, 44.5% mIoU on COCO Object dataset, which can be adopted as a strong baseline for further research.

Ablation study on relation-aware matching. As described in Sec. 3.5, we propose to model the beneficial structure information contained in inter-class relationships via relation-aware matching. Compared to the naive approach that only considers only the cosine similarity between region features and each individual category reference feature, our proposed strategy achieves a sizeable gain, *i.e.*, 1.3% in mIoU, as presented in the 3^{rd} and 4^{th} rows of Table 2. The results prove that the proposed similarity measurement based on ranking distribution facilitates effective disambiguation in the matching process, which brings a notable reduction in the misclassification of regions.

After integrating subcategory reference features into relation-aware matching, the performance further improved from 43.1% to 44.5% in mIoU, as shown in 5^{th} row of Table 2. This improvement can be attributed to the ability of the SD model to generate diverse image features, enabling our method to effectively handle intra-class diversity. Moreover, the scoring scheme in Equ. 10, akin to a voting mechanism, further enhances the robustness of the matching process. Furthermore, in Table 5, we demonstrate that reference features

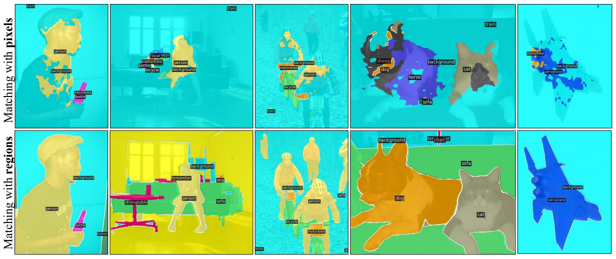

Figure 4. Effectiveness of SAM based region-level matching. The SAM could well capture visual concepts within images.

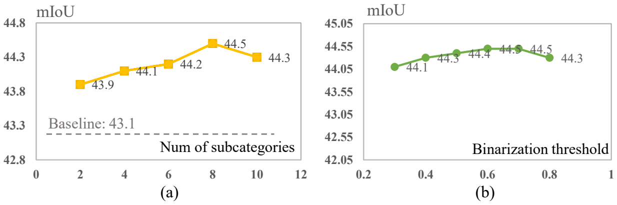

Figure 5. Hyperparameter experiments on the number of subcategories and binarization threshold of cross-attention map.

that share high similarities with region features are better candidates to serve as category agents.

Ablation on DINOv2. To demonstrate the superiority of feature extraction using DINOv2, we conduct comparison experiments of CLIP [38], MAE [17], Stable Diffusion Unet (SD) [42], and DINOv2 as presented in Table 3. DINOv2 achieves the best performance attribute to effective pretraining based on image-level and patch-level discriminative self-supervised learning. While CLIP fails to match region-level features precisely, Stabel Diffusion UNet is constrained to some extent under the unconditional settings. MAE exhibits suboptimal performance in our setting due to its relatively limited high-level semantic modeling capability.

Ablation on SAM. Our relation-aware matching operates at the region level, as SAM furnishes comprehensive mask proposals for the test images. To explore the effectiveness of SAM, on the PASCAL VOC dataset, we compare SAM with modified MaskFormer [9] pre-trained on the COCO-stuff dataset following [24]. We also construct a vanilla baseline that compares category reference features with pixel features of test images directly. As shown in Table 4, both segmenters outperform the baseline by a large margin. We deem the main reason is that the region-level matching is more robust than pixel-level one. As illustrated in Fig. 4, the pixel classification paradigm fails to capture all visual concepts within images. SAM slightly surpasses MaskFormer as SAM can generate higher-quality masks on novel classes, naturally more suitable for open-vocabulary tasks.

Investigation of the intra-modal matching. In Fig. 6, we visualized the t-SNE plots of the region features alongside their corresponding text-modal reference features and image-modal reference features using the same image based on CLIP [38] and DINOv2 [35], respectively. It can be observed that in the all-purpose feature space of DINOv2, the region features of different categories are well-differentiated and align well with their corresponding reference features. However, in the case of CLIP, different region features not only exhibit notable overlap but also fail to align effectively with corresponding text-modal reference features. This difference substantiates the underlying rationale of our motivation, *i.e.*, intra-modal matching.

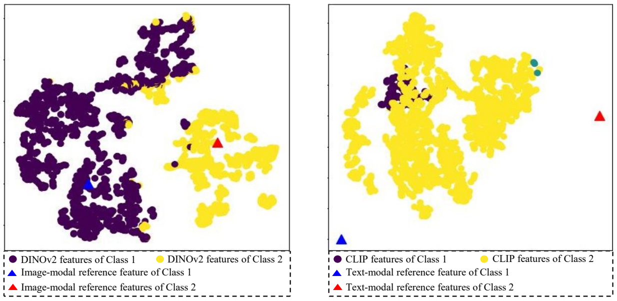

Figure 6. T-SNE visualization of features of different region, as well as corresponding intra-modal and cross-modal reference features.

Hyperparameter Evaluations. Quantitative experiments are conducted to find a suitable number of subcategories used in relation-aware matching. As illustrated in Fig. 5 (a), the performance continues to grow until the number achieves 8, beyond which it starts to decline. It is reasonable as an appropriate number of subcategory references can model category diversity, but over-partitioning may lead to deviations from instance features. We then explore how the binarization threshold affects the performance in Fig. 5 (b). The model shows low sensitivity to threshold values owing to the capability of SAM, with a setting of 0.7 yields marginally better results.

5. Conclusion

In this work, we presented a training-free Relation-aware Intra-modal Matching (RIM) framework to tackle the challenging open-vocabulary semantic segmentation. We construct better-aligned image-modal category reference features based on the Stable Diffusion model and SAM. Then a relation-aware matching strategy is employed for region classification. RIM not only achieves results significantly surpassing the state-of-the-art but also opens up new avenues for OVS from an image-to-image matching perspective.

6. Acknowledgments

This work was partially supported by the National Defense Basic Scientific Research Program of China (Grant JCKY2021130B016), National Nature Science Foundation of China (Grant 12150007), Youth Innovation Promotion Association CAS 2018166.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[4] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 6

[5] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 699–710, 2023. 6

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021. 1, 2, 3

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 8

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[11] 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. 1

[12] 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. *arXiv preprint arXiv:2010.11929*, 2020. 5

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

[15] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1086–1096, 2023. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 6, 8

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[19] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7014–7023, 2018. 3

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[21] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 6

[22] 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. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3, 5

[23] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8334–8343, 2021. 3

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1, 2, 3, 8

[25] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3, 5, 6

[26] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 6

[27] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023. 3

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[29] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 1, 3, 6

[30] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17927, 2023. 2

[31] Naisong Luo, Rui Sun, Yuwen Pan, Tianzhu Zhang, and Feng Wu. Electron microscopy images as set of fragments for mitochondrial segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2

[32] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19617–19626, 2023. 3

[33] Huayu Mai, Rui Sun, Yuan Wang, Tianzhu Zhang, and Feng Wu. Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 3

[34] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 5

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4, 5, 8

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 3

[37] Yuwen Pan, Naisong Luo, Rui Sun, Meng Meng, Tianzhu Zhang, Zhiwei Xiong, and Yongdong Zhang. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21474–21484, 2023. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6, 7, 8

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *URL https://arxiv.org/abs/2204.06125*, 7, 2022. 3

[40] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in vision-language models. *arXiv preprint arXiv:2210.09996*, 2022. 6

[41] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023. 6

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 5, 6, 8

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 3

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[45] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 2

[46] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 2

[47] Rui Sun, Naisong Luo, Yuan Wang, Yuwen Pan, Huayu Mai, Zhe Zhang, and Tianzhu Zhang. 1st place solution for youtubevos challenge 2022: Video object segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 2

[48] Rui Sun, Naisong Luo, Yuwen Pan, Huayu Mai, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Appearance prompt vision transformer for connectome reconstruction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1423–1431. International Joint Conferences on Artificial Intelligence Organization, 2023. Main Track. 2

[49] Rui Sun, Huayu Mai, Naisong Luo, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2023. 2

[50] Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, 2023. 3

[51] Rui Sun, Yuan Wang, Huayu Mai, Tianzhu Zhang, and Feng Wu. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1218–1228, 2023. 2

[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 6

[53] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 6

[54] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12275–12284, 2020. 3

[55] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022. 3

[56] Yuan Wang, Naisong Luo, and Tianzhu Zhang. Focus on query: Adversarial mining transformer for few-shot segmentation. In *Advances in Neural Information Processing Systems*, 2023.

[57] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2023. 3

[58] Li Wangkai, Li Zhaoyang, Sun Rui, Mai Huayu, Luo Naisong, Yuan Wang, Pan Yuwen, Xiong Guoxin, Lai Huakai, Xiong Zhiwei, et al. Maunet: Modality-aware anti-ambiguity u-net for multi-modality cell segmentation. In *Competitions in Neural Information Processing Systems*, pages 1–12. PMLR, 2023. 2

[59] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *arXiv preprint arXiv:2308.06160*, 2023. 2

[60] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 2, 3, 4

[61] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 3

[62] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 2

[63] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[64] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1, 3, 6

[65] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023. 3, 6

[66] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 3

[67] 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 *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 1, 2, 3

[68] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 1, 3

[69] Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7071–7080, 2023. 2, 3, 6

[70] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*, 2023. 3

[71] Yifan Zhang, Bo Pang, and Cewu Lu. Semantic segmentation by early region proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1258–1268, 2022. 2

[72] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[73] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 2

[74] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 1, 3

[75] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 6