IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 May 2024, accepted 19 June 2024, date of publication 24 June 2024, date of current version 1 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3418210

RESEARCH ARTICLE

TAG: Guidance-Free Open-Vocabulary Semantic Segmentation

YASUFUMI KAWANO AND YOSHIMITSU AOKI, (Member, IEEE)
Graduate School of Integrated Design Engineering, Keio University, Kanagawa 223-8522, Japan
Corresponding author: Yasufumi Kawano (ykawano@aoki-medialab.jp)

ABSTRACT Semantic segmentation is a crucial task in computer vision, where each pixel in an image is classified into a category. However, traditional methods face significant challenges, including the need for pixel-level annotations and extensive training. Furthermore, because supervised learning uses a limited set of predefined categories, models typically struggle with rare classes and cannot recognize new ones. Unsupervised and open-vocabulary segmentation, proposed to tackle these issues, faces challenges, including the inability to assign specific class labels to clusters and the necessity of user-provided text queries for guidance. In this context, we propose a novel approach, **TAG** which achieves **T**raining, **A**nnotation, and **G**uidance-free open-vocabulary semantic segmentation. TAG utilizes pre-trained models such as CLIP and DINO to segment images into meaningful categories without additional training or dense annotations. It retrieves class labels from an external database, providing flexibility to adapt to new scenarios. Our TAG achieves state-of-the-art results on PascalVOC, PascalContext and ADE20K for open-vocabulary segmentation without given class names, i.e. improvement of +15.3 mIoU on PascalVOC.

INDEX TERMS Semantic segmentation, open-vocabulary, zero-guidance.

I. INTRODUCTION

Semantic segmentation represents a crucial task in computer vision, which describes assigning class labels to each pixel of an image. Its applications span diverse domains, including robotics and satellite image analysis.

Despite its significance, current semantic segmentation methods still face several critical challenges. Firstly, these methods are high-cost, requiring pixel-level annotation and extensive training. Secondly, since supervised learning depends on a predefined set of categories, detecting extremely rare or completely new classes during prediction is virtually impossible.

Two related tasks were proposed to address these limitations: unsupervised and open-vocabulary semantic segmentation. Unsupervised semantic segmentation [1], [2], [3] avoids the expensive annotation process by using representations obtained through a backbone model [4], [5] trained on a different task. Open-vocabulary semantic segmentation [6], [7], [8], [9], [10] enables the identification of a wide array of categories through natural language and is not bound to a pre-defined set of categories.

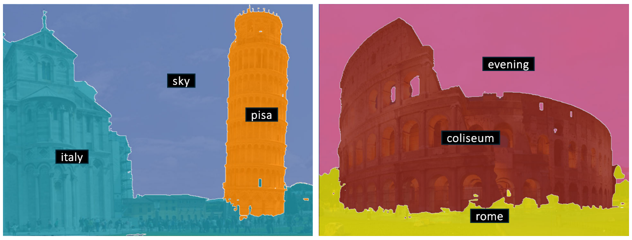

FIGURE 1. Guidance-free Open-Vocabulary Semantic Segmentation. Our TAG can segment an image into meaningful segments without training, annotation, or guidance. It successfully segments structures such as the Leaning Tower of Pisa and the Colosseum. Unlike traditional open-vocabulary semantic segmentation methods, TAG can segment and categorize without text-guidance.

However, there are still challenges to be solved with these methods. Unsupervised semantic segmentation clusters images by class but cannot identify the class of each cluster, while open-vocabulary segmentation assumes that text queries describing objects in the image are provided by the user. To address these challenges, zero-guidance segmentation emerged in [14], enabling open-vocabulary

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

TABLE 1. Relationship with related works. We categorize related works into four distinct areas: training-free, (dense) annotation-free, guidance-free, and open-vocabulary.

Method	Pretrained Backbone	Training Free	Annotation Free	Guidance Free	Open Vocabulary
Traditional Semantic Segmentation					
DeepLab [11]	-	-	-	-	-
Unsupervised Segmentation					
STEGO [2]	DINO [4]	-	√	√	-
HP [3]	DINOv2 [5]	-	√	√	-
Open Vocabulary Segmentation					
ODISE [8]	StableDiffusion [12]	-	-	-	√
MaskCLIP [9]	CLIP [13]	√	√	-	√
Zero-Guidance Segmentation					
ZeroSeg [14]	CLIP [13]& DINO [4]& GPT-2 [15]	√	√	√	√
SelfSeg [16]	BLIP [17] & X-Decoder [18]	√	-	√	√
TAG (Ours)	CLIP [13] & DINOv2 [5]	√	√	√	√

segmentation without the need for inputting class candidates (guidance), yet there is still room for improvement in terms of performance. We categorize these related works into four distinct areas in Table 1.

Based on these backgrounds, we further improved this approach by introducing a novel method named **TAG**, which offers higher performance and flexibility. As its primary strength, TAG achieves **T**raining, **A**nnotation, and **G**uidance-free open-vocabulary semantic segmentation. This method employs a novel approach by extracting semantic features from each pixel in an image using CLIP [13], and then retrieving the open-vocabulary classes based on these features from an external database [19], [20], [21], [22]. TAG operates using pre-trained frozen models CLIP [13] and DINOv2 [5], eliminating the need for an additional training process. CLIP [13] can identify diverse objects and scenes while its segmentation results are often coarse and noisy, necessitating refinement. We use DINOv2 [5] to excel in capturing fine details and global context, enabling precise segmentation. Combining these models leverages CLIP [13]'s generalization and DINO [5]'s detailed feature extraction for more accurate segmentation. These models do not utilize the dense and costly annotations traditionally required for semantic segmentation.

Furthermore, through the extensibility of its database, this method also incorporates flexibility, making it easy to adapt to new classes or scenarios. A major distinction between previous methods [14], [16], and our TAG is that it provides more flexibility as it can be extended to include new concepts by adding them to the database while previous methods require re-training. It is important to note that while the database used in TAG is finite, the language models like BLIP [17] or GPT [15] are also constructed from similarly finite datasets. In [23], it is even reported the retrieval-based methods provided superior results over BLIP [17] in the context of image classification.

Our TAG can segment an image into meaningful segments as shown in Figure 1 without any text guidance. In particular, TAG is able to accurately segment structures with their proper nouns, such as the *Leaning Tower of Pisa* and the *Coliseum*. In addition, TAG shows significant improvements in contrast to other comparable segmentation methods, i.e. on the PascalVOC [24] dataset (+15.3 mIoU).

Our contributions are the following:

1) We propose a novel approach, namely TAG, to achieve open-vocabulary semantic segmentation that does not require pre-defined categories by retrieving segment categories from an external database.
2) TAG achieves compelling segmentation results for all categories in the wild without any additional training, high-cost dense annotation, or text query guidance.
3) TAG outperforms the previous state-of-the-art methods by 15.3 mIoU on the PascalVOC [24] dataset, demonstrating the superior segmentation performance of our proposed approach.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Semantic segmentation is the task of assigning class labels to all pixels in an image, commonly using convolutional neural networks [11], [25] or vision transformers [26] for end-to-end training. These methods, while effective, depend on extensive annotation and significant computational resources for training, and are limited to predefined categories. Thus, unsupervised, and domain-flexible approaches have recently gained importance.

Unsupervised semantic segmentation [1], [2], [3], [27] attempts to solve semantic segmentation without using any kind of supervision. STEGO [2] and HP [3] optimize the head of a segmentation model using image features obtained from a backbone pre-trained by DINO [4] and DINOv2 [5], an unsupervised method for many tasks. However, unsupervised semantic segmentation clusters images by class but cannot identify the class of the each cluster. In contrast, our TAG distinguishes classes without extra training or annotation.

B. OPEN-VOCABULARY SEMANTIC SEGMENTATION

Open vocabulary semantic segmentation, crucial for segmenting objects across domains without being limited to predefined categories, has seen notable advancements with the introduction of key methodologies [6], [7], [7], [8], [9], [10], [28], [29], [30], [31], [32], [33], [34].

Early attempts, such as ZS3Net [30] and SPNet [32], focused on zero-shot learning, training custom modules to bridge visual and language embedding spaces. These methods set the foundation for future improvements.

This area has seen significant improvement, particularly through integrating vision-language models like CLIP [13], which train visual and textual feature encoders on extensive image-text pairs. LSeg [33], OpenSeg [7], OPSNet [34], and OVSeg [7] have each contributed to the advancements in the field leveraging CLIP [13]. These methods typically generate class-agnostic masks before using CLIP [13] to classify each mask, demonstrating the versatility of CLIP [13] embeddings in open vocabulary semantic segmentation.

Moreover, MaskCLIP [9] and GEM [35] have highlighted the potential of using intermediate representations from a frozen CLIP [13] encoder to directly segment images without additional training, reducing both annotation and training costs. Concurrently, models like ODISE [8] have explored the integration of pre-trained diffusion models [12] with CLIP [13] to extend to high performance panoptic segmentation.

Despite these advancements, a limitation across these methods is their reliance on text input as guidance from users. Our TAG tackles this limitation and allows for open-vocabulary segmentation without text guidance. Closest and concurrent to our work is the zero-guidance semantic segmentation paradigm [14], in which clustered DINO [4] embeddings are combined with CLIP [13]. To generate captions from CLIP [13] features, ZeroSeg [14] uses ZeroCap [36] which combines a language model, GPT-2 [15], with CLIP [13]. It adjusts parts of GPT-2 [15] to finish the sentence, starting with "Image of a ..." so that the sentence closely matches the images according to CLIP's understanding.

However, there is still room for improvement in terms of performance. We hypothesize that the issue is related to the performance of ZeroCap [36]. Therefore, as a new method, our TAG uses a novel approach that retrieves categories from a database for estimating categories.

C. TEXT RETRIEVAL FROM CLIP EMBEDDING

In natural language processing, retrieving information from external databases has been shown to boost the performance of large language models [37], [38], [39]. This concept is also explored in computer vision, particularly for addressing class imbalance by using databases to retrieve training samples or image-text pairs. RAC [40] and VIC [23] achieved image classification without relying on predefined classes by utilizing an external database. It has the advantage of low memory consumption because it only uses captions from databases like Public Multimodal Datasets (PMD) [19] collecting image-text pairs from different public datasets.

III. METHOD

Figure 2 shows an overview of our proposed method which we call TAG, a novel approach. Our TAG attempts to partition input images into semantic segments and label each segment with open-vocabulary categories. To this end, we propose to identify segment candidates using per-pixel features obtained from DINOv2 [5] (Sec. III-A), acquire representative segment embeddings for segment candidates using per-pixel features from a ViT pre-trained with CLIP [13] (Sec. III-B), and assign categories to each candidate segment by retrieving the closest matching sentence from an external database (Sec. III-C). Note that, unlike traditional open-vocabulary semantic segmentation, the input is only the image, with no need to input category candidates as guidance.

A. SEGMENT CANDIDATES WITH DINO

It has been observed that segmentation results obtained from CLIP-based segmentation methods [9], [35] are fragmented and noisy as shown in Figure 4. Therefore, the first step in our TAG pipeline is calculating segmentation candidates to achieve more accurate segmentation results. To obtain more precise segmentation outcomes than CLIP-based methods without using dense annotations, we reference unsupervised segmentation methods [2], [3] and employ a ViT pre-trained with DINOv2 [5].

The output of DINOv2 [5] is a feature map $\in \mathbb{R}^{D \times \frac{H}{P} \times \frac{W}{P}}$, where D is the dimension of the feature, P is the patch size of the transformer, and H and W are image sizes. This feature map will be upsampled to per-pixel features $\in \mathbb{R}^{D \times H \times W}$. Once the per-pixel feature is obtained, to assign categories for each segmentation candidate, we use k-means to divide the per-pixel features into segmentation candidates, resulting in oversegmentation.

B. REPRESENTATIVE SEGMENT EMBEDDINGS WITH CLIP

CLIP [13] is a ViT model that can embed images and text into the same latent space. To assign natural language categories to each segment, we use CLIP [13] to embed the image at the pixel level.

Instead of directly acquiring pixel-level features from CLIP [13], we extract dense patch-level features from the image encoder of CLIP [13] following CLIP-based segmentation methods [9], [35]. The image encoder of CLIP [13] uses a multi-head attention layer, where the globally average-pooled feature works as the query, and the feature at each patch generates a key-value pair. Then, this layer outputs a spatial weighted sum of the incoming feature map followed by a linear layer $F(\cdot)$:

$$\begin{aligned}\mathrm{AttnPool}(\overline{q}, k, v) &= F\left(\sum_i \mathrm{softmax}\left(\frac{\overline{q}k_i^T}{C}\right) v_i\right) \\ &= \sum_i \mathrm{softmax}\left(\frac{\overline{q}k_i^T}{C}\right) F(v_i), \qquad (1)\end{aligned}$$

$$\overline{q} = \mathrm{Emb}_q(\overline{x}),\ k_i = \mathrm{Emb}_k(x_i),\ v_i = \mathrm{Emb}_v(x_i), \qquad (2)$$

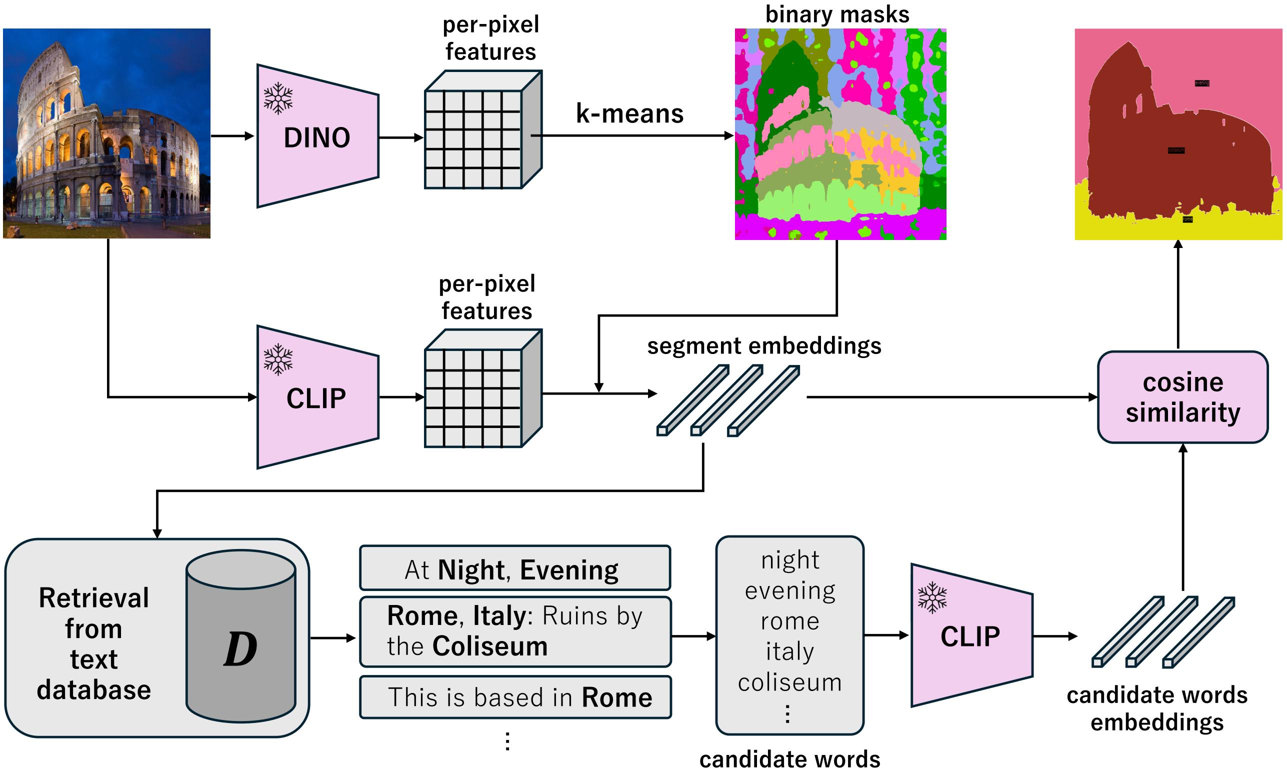

FIGURE 2. High-level overview of our TAG architecture. Our TAG can partition images into semantic segments and label each segment with open-vocabulary categories. First, TAG identifies segment candidates using per-pixel features obtained from DINOv2 [5]. Then, it acquires representative segment embeddings for segment candidates using per-pixel features from a ViT pre-trained with CLIP [13]. Finally, the categories are assigned to each candidate segment by retrieving the closest matching sentence from an external database. Note that the input is only the image, with no need to input category candidates as guidance.

where C denotes a constant scaling factor and Emb$(\cdot)$ represents a linear embedding layer. x_i is the input feature at patch i and $\overline{x}$ is the average of all x_i. The Transformer layer in CLIP [13] outputs a detailed image representation, made possible because $F(v_i)$, computed at each spatial location, captures a rich response of local semantics.

Based on this observation, we utilize the features from the last attention layer of CLIP [13] image encoder by adopting the GEM [35] mechanism.

CLIP model in TAG outputs value features $\in \mathbb{R}^{D\times\frac{H}{P}\times\frac{W}{P}}$, where D is the dimension of the feature, P is the patch size of the transformer. These features contain dense representations of the image, capturing patch-level information, which we upsample to per-pixel features $\in \mathbb{R}^{D\times H\times W}$, corresponding to the same size as the features obtained from DINO [5].

Next, to assign categories to segment candidates, we calculate embedding features representing the segments from CLIP [13] per-pixel features $\mathbf{f} \in \mathbb{R}^{D\times H\times W}$ as shown in Figure 3. For each segment k, this representative segment embedding $\bar{\mathbf{f}}_k \in \mathbb{R}^D$ is computed by averaging based on the values $m_{khw} \in \{0,1\}$, which results from applying k-means to the output of the DINO [5] with k classes, as follows:

$$\bar{\mathbf{f}}_k = \frac{1}{M_k}\sum_{h,w} m_{khw}\cdot \mathbf{f}_{hw}, \quad M_k = \sum_{h,w} m_{khw} \tag{3}$$

C. SEGMENT CATEGORY RETRIEVAL

CLIP [13] can embed images and text in the same latent space, but the model itself cannot generate images or text from the embedded features. To address this challenge, our proposed method TAG finds the closest category using multi-modal data from large databases.

First, we retrieve a few of the most probable candidate classes from the large classification space.

Let D be the database of image captions. Given representative segment embedding $\bar{\mathbf{f}}_k$, retrieve the set $D_{\bar{\mathbf{f}}_k} \subset D$ of the n closest captions to each segment embedding by

$$D_{\bar{\mathbf{f}}_k} = \underset{\mathbf{d}\in D}{\text{top-}n}\ \frac{\bar{\mathbf{f}}_k^T\cdot \mathbf{f}_d}{\|\bar{\mathbf{f}}_k\|\cdot\|\mathbf{f}_d\|}, \quad \mathbf{f}_d = \text{CLIP}_t(d), \tag{4}$$

where CLIP$_t$ is the text encoder of the CLIP [13].

Next, to extract candidate words $C_{\bar{\mathbf{f}}_k}$ from the set $D_{\bar{\mathbf{f}}_k}$, we create a set of all words that are contained in the captions. We sequentially apply three operations: (i) remove noisy candidates, (ii) standardize their format, and (iii) filter them.

In the first operation, we remove all the irrelevant words, such as URLs, or file extensions.

Secondly, we align words referring to the same semantic class in a standardized format. Specifically, Converting upper case to lower case and plural words to singular format.

In the final operation, we filter out two types of words: rare and noisy categories based on the frequency of word

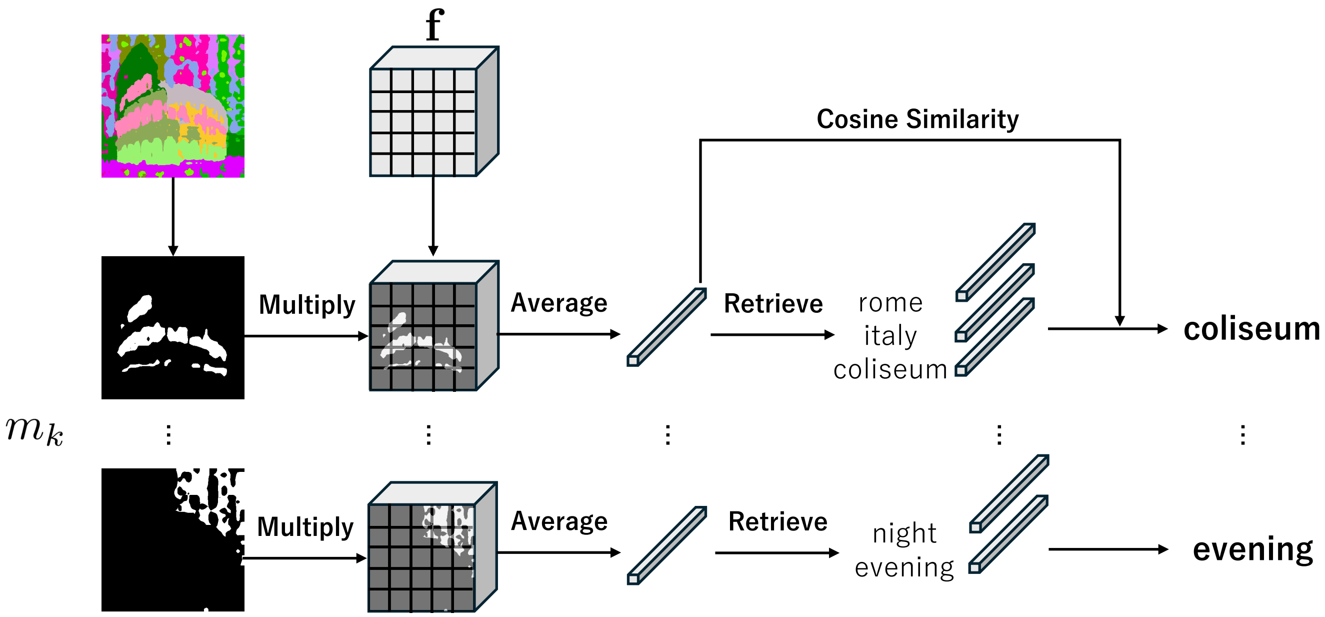

FIGURE 3. Overview of the flow for each segment. Each segment independently retrieves for category candidates and assigns a category.

occurrences, as well as entire categories of words determined by Part-Of-Speech (POS) [41] tagging. Frequency filtering involves retaining only those words that appear more than two times in the input text. If the threshold is set too high and no words meet the criterion, it is lowered to include at least the most frequently occurring words. The POS [41] tagging classifies words into groups like adjectives, articles, nouns, or verbs, allowing us to exclude any terms that do not hold semantic significance as segmentation categories.

Given candidate words $C_{\bar{\mathbf{f}}_k}$, we assign words to representative segment embedding $\bar{\mathbf{f}}_k$ as

$$W = \operatorname*{argmax}_{\mathbf{c} \in C_{\bar{\mathbf{f}}_k^T}} \frac{\bar{\mathbf{f}}_k^T \cdot \mathbf{f}_c}{\|\bar{\mathbf{f}}_k\| \cdot \|\mathbf{f}_c\|}, \quad \mathbf{f}_c = \mathrm{CLIP}_t(c), \tag{5}$$

where W is assigned a category word to segment. Through the above process, we can obtain segmentation results by assigning categories to each segment candidate.

IV. EXPERIMENT

First, we present the implementation details in Section IV-A. Next, we compare our results to previous methods in Section IV-B and evaluate the open vocabulary aspect in Section IV-C. Finally, we justify the construction of TAG through an ablation study in Section IV-D.

A. IMPLEMENTATION DETAILS

For our implementation of TAG, we employed a frozen pre-trained CLIP [13] and DINO [5] with ViT-L/14 architecture and input 448 × 448 images to them. As database, we use PMD [19], CC12M [20], WordNet [22] and English-Words [21]. In addition, we use a fast indexing technique, FAISS [45]. Our model works with a GPU memory of 15 GB.

B. MAIN RESULTS

To validate the performance of TAG, we conducted comprehensive comparative experiments with its closest counterpart, ZeroSeg [14]. For settings, TAG uses a PMD database [19]. We set the number of k-means clusters as 15 and the frequency filtering threshold 2. For the evaluation, we used the mean Intersection over Union (mIoU) as the primary metric. The predicted text T_i needs to be assigned to one of the ground truth classes T^{gt}. T_i is assigned to the ground-truth label that is closest in the Sentence-BERT [46] embedding space, following the same approach as ZeroSeg [14]. Formally, the new label T_i^* is computed by

$$T_i^* = \operatorname*{argmax}_{t \in T^{gt}} [\mathrm{cossim}^{\mathrm{SBERT}}(T_i, t)]. \tag{6}$$

We perform our experiments on the PascalVOC [24] dataset comprising 20 classes, PascalContext [43] with 59 classes, as well as ADE20K [44] consisting of 150 classes.

The qualitative results are shown in Figure 4 and Figure 5. In Figure 4, we compared TAG with CLIP [13] base open-vocabulary method, MaskCLIP [9], and GEM [35]. Using MaskCLIP [9] and GEM [35] results in a noisy and fragmented segmentation, whereas TAG achieves more consistent segments that better correspond with the shape of the object and segment categories. In Figure 5, we compare GroupViT [42], ZeroSeg [14], and our TAG on images containing general objects from PascalContext [43]. In the image (a), TAG is the only method that accurately recognizes a cow as a calf. In addition, TAG assigns the precise and relevant class ‘barn’ to the surroundings of the calf, unlike ZeroSeg which incorrectly includes the class ‘sleeping’. In image (b), TAG is the only method to correctly identify ‘sunglasses’, and it also accurately classifies the dog as a ‘bulldog’. However, in image (c), TAG does not distinguish

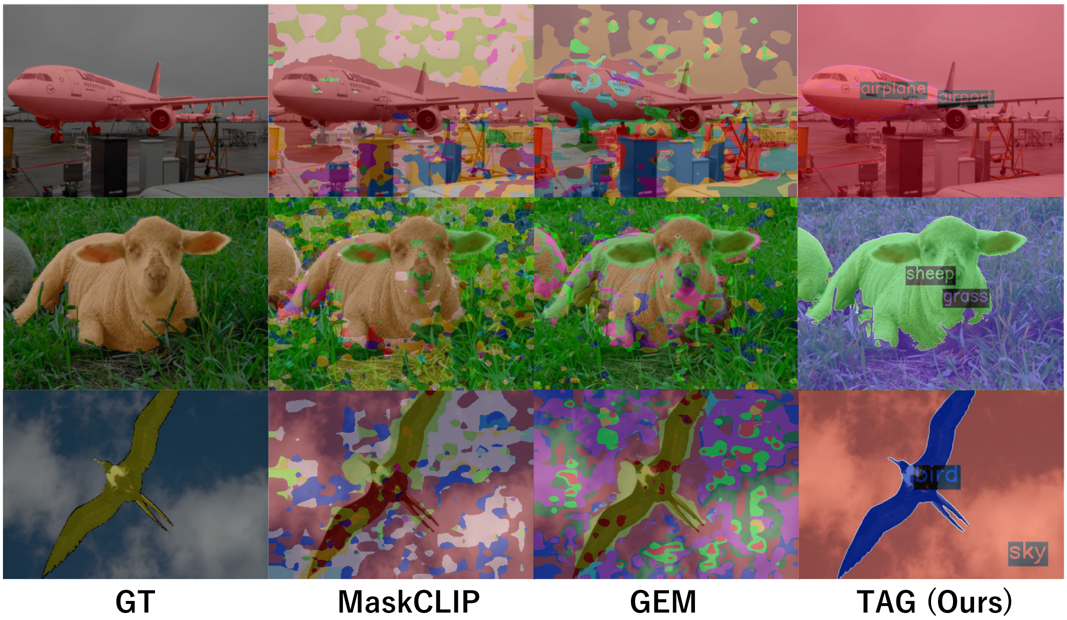

FIGURE 4. Comparison results with CLIP base open-vocabulary segmentation methods on PascalVOC [24] Note that MaskCLIP [9] and GEM [35] uses text guidance while our TAG does not use.

TABLE 2. Comparison of state-of-the-art methods. We evaluate on PascalVOC [24], PascalContext [43], and ADE20K [44] and report the mIoU. The clear boost in performance by retrieving open-vocabulary segment categories underlines their semantic richness and effectiveness.

Method	Training Free	PascalVOC [24]	PascalContext [43]	ADE20K [44]
		20class	59class	150class
Open-Vocabulary Segmentation				
X-Decoder [18]	-	96.2	69.2	6.4
ODISE [8]	-	82.7	55.3	11.0
OpenSeg [7]	-	72.2	44.8	8.8
GroupViT [42]	-	50.7	25.9	-
LSeg [33]	-	47.4	-	-
GEM [35]	√	26.5	11.8	4.4
MaskCLIP [9]	√	28.6	25.5	-
Zero-Guidance Segmentation				
ZeroSeg [14]	√	20.1	19.6	-
SelfSeg [16]	√	41.6	-	6.4
TAG (Ours)	√	**56.9**	**20.2**	**6.6**

between a desk and a chair but rather assigns the rough class 'room' to the entire space. Occasionally, as shown in (d), TAG assigns proper nouns such as 'swindon' and 'swanage' which are names of cities in South England. While TAG correctly identifies the background as the city 'swanage', the ground is incorrectly assigned to the city of 'swindon'. We hypothesize this is caused by both segments being close in the CLIP [13] embedding space.

The quantitative results are shown in Table 2. TAG shows an improvement of +15.3 mIoU on PascalVOC [24], +0.6 mIoU on PascalContext [43], and +0.2 mIoU on ADE20K [44] compared to previous zero-guidance segmentation state-of-the-art results. In particular, TAG shows a dramatic performance improvement on PascalVOC [24], which was identified as a limitation in ZeroSeg [14]. Additionally, our TAG method has made significant improvements compared to untrained open-vocabulary segmentation methods, demonstrating an impressive improvement of +28.3 mIoU on PascalVOC [24], even without text-based guidance.

C. OPEN VOCABULARY SEGMENTATION ON WEB-CRAWLED IMAGES

In this section, we thoroughly assess the performance of TAG using open vocabulary segmentation experiments on web-crawled images, where we test the model's ability to accurately segment various unseen classes, including specific and detailed categories such as 'joker' and 'porsche'.

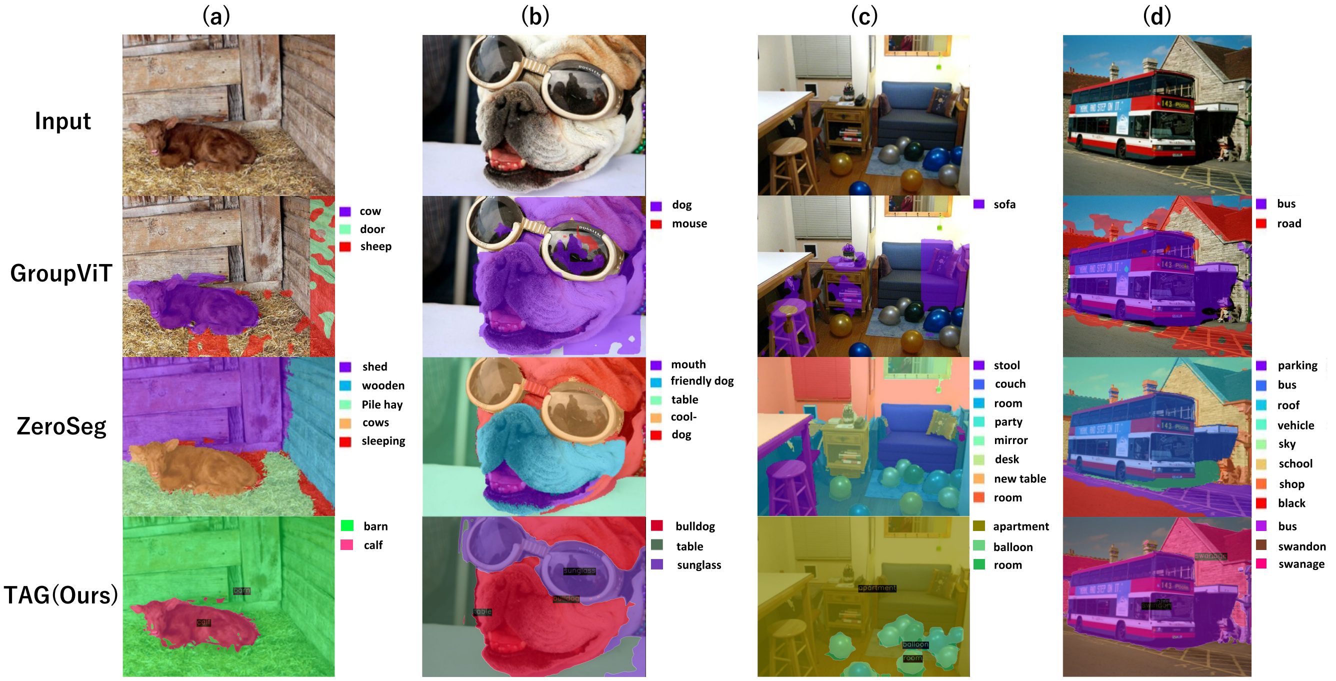

FIGURE 5. Qualitative results. We compare GroupViT [42], ZeroSeg [14], and our TAG on images containing general objects from PascalContext [43]. This figure indicates that TAG can segment and label correctly.

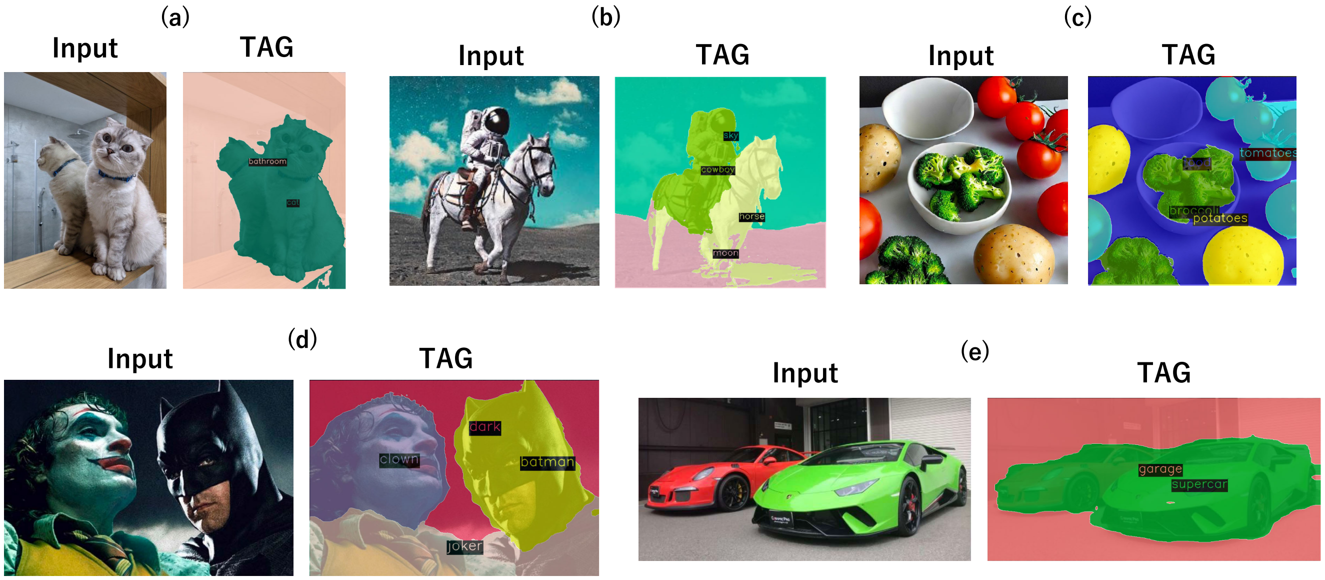

FIGURE 6. Open-vocabulary segmentation results. In (a) we test on a general image, (b) and (c) show images generated by Stable Diffusion [12], and (d) and (e) are images featuring specific proper nouns.

The qualitative outcomes of the experiments are visually depicted in Figure 6. In this figure, (a) represents a general image, while (b) and (c) showcase images created by Stable Diffusion [12]. Furthermore, (d) and (e) show images containing proper nouns.

In image (a), although the complex concept of a 'mirror' is not captured, the segmentation successfully identifies both 'cat' and 'bathroom', resulting in accurate outcomes. In image (b), while failing to recognize the 'astronaut', the model aptly estimates the ground as the 'moon', leading to a logical result. Given TAG's ability to identify the ground as the moon, it is evident that it can understand the whole image while generating segment embeddings. Image (c) showcases the precise segmentation of various foods. Image (d) impressively segments and identifies proper nouns such as 'joker' and 'batman', demonstrating remarkable results. Lastly, image (e), despite containing the specific proper noun 'porsche', is correctly recognized as a supercar, affirming the accuracy of the segmentation.

TABLE 3. Ablation study on database and numbers of cluster used for k-means. We evaluate on PascalVOC [24], PascalContext [43], and ADE20K [44] and report the mIoU. These results reveal that using PMD [19] and CC12M [20] as the database and setting the number of k-means clusters to 15 is the most robust choice, consistently yielding favorable outcomes across multiple datasets.

Database	Cluster Numbers	PascalVOC [24]	PascalContext [44]	ADE20K [45]
		20class	59class	150class
PMD [19]	5	57.5	19.2	5.0
PMD [19]	10	56.9	19.1	6.2
PMD [19]	15	56.9	**20.2**	**6.6**
PMD [19]	20	56.4	18.9	5.2
PMD [19]	25	51.8	19.5	5.2
CC12M [20]	5	**58.5**	19.2	4.9
CC12M [20]	10	58.4	19.2	6.1
CC12M [20]	15	58.1	19.6	6.1
CC12M [20]	20	57.6	19.0	5.0
CC12M [20]	25	52.9	19.1	5.1
WordNet [22]	15	54.3	17.5	4.5
EnglishWords [21]	15	43.7	14.5	3.0

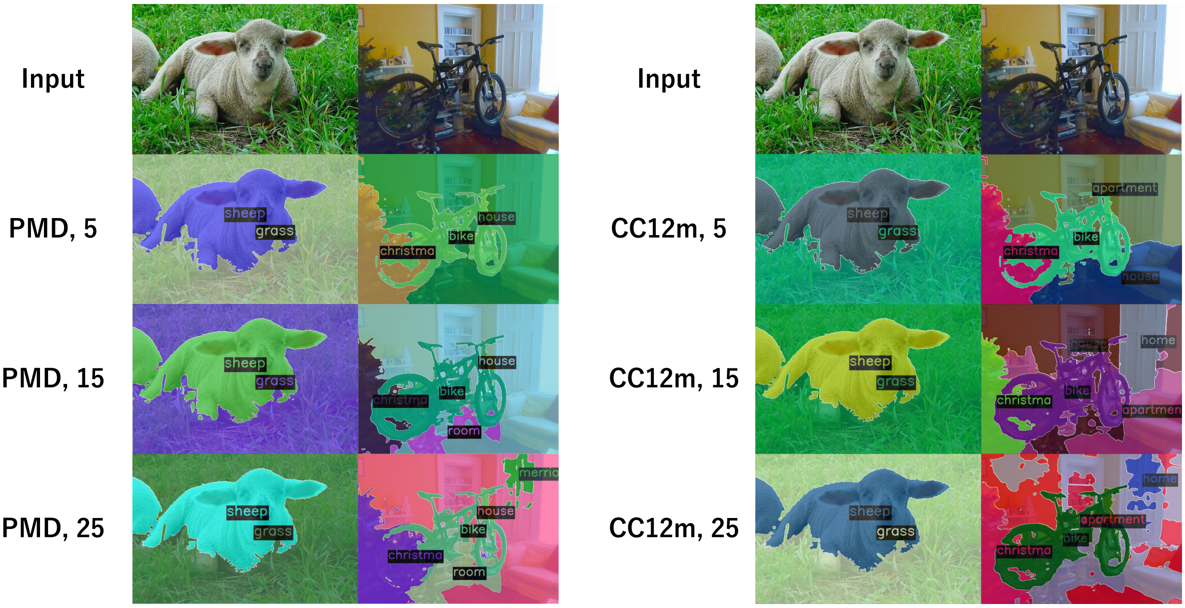

FIGURE 7. Qualitative results of ablation study on PascalVOC [24]. The database and the number of k-means clusters are shown with the results.

These findings serve as compelling evidence that TAG exhibits robust capabilities to accurately segmenting open vocabularies, including complex and specific categories, thus underscoring its versatility and effectiveness in handling diverse and intricate segmentation tasks.

D. ABLATION STUDY

In this section, we perform ablation studies on TAG, examining how various databases, cluster numbers of k-means, and label reassignment of evaluation affect performance.

Table 3 presents the results of the ablation experiments comparing the effect of databases and cluster numbers of k-means on the mIoU. The results indicate that PMD [19] and CC12M [20] are preferable datasets for our database. These results also reveal that using PMD [19] as the database and setting the number of k-means clusters to 5 is the most robust choice, consistently yielding favorable outcomes across multiple datasets. Figure 7 shows the ablation qualitative results. Left image remains unchanged regardless of variations in the database or the number of clusters, while increasing the number of clusters has been observed to cause segments with the same semantic meaning, such as 'apartment' to be divided into different segments like 'home' or 'house' in right image.

Furthermore, we conduct ablation experiments on filtering operations and the number of captions to select used for segment category retrieval. Table 4 shows that utilizing all three filtering operations yields the best results. Similarly, we examine the effect of the threshold on the frequency filtering. The results are shown in Table 5 and indicate that using our default threshold of 2 is justified. In addition, Table 6 shows the number of captions to select and indicates that our default setting threshold to 10 is appropriate..

TABLE 4. Ablation study on filtering operations. We evaluate on PascalVOC [24], PascalContext [43], and ADE20K [44] and report the mIoU. This table shows that utilizing all three filtering operations yields the best results.

Remove	Standardize	Filter	PascalVOC [24]	PascalContext [43]	ADE20K [44]
			20class	59class	150class
			54.5	18.4	5.0
√			55.2	18.5	5.1
√	√		56.0	18.6	5.8
√	√	√	**56.9**	**20.2**	**6.6**

TABLE 5. Ablation study on the threshold of frequency filtering. We evaluate on PascalVOC [24], PascalContext [43], and ADE20K [44] and report the mIoU.

threshold	PascalVOC [24]	PascalContext [43]	ADE20K [44]
	20class	59class	150class
1	54.3	18.6	5.8
2	**56.9**	**20.2**	6.6
5	54.1	18.6	**6.8**
10	51.7	17.8	6.3
20	51.4	17.8	5.8

TABLE 6. Ablation study on the number of retrieved captions. We evaluate on PascalVOC [24], PascalContext [43], and ADE20K [44] and report the mIoU.

Captions Number	PascalVOC [24]	PascalContext [43]	ADE20K [44]
	20class	59class	150class
5	55.9	19.5	6.3
10	56.9	**20.2**	**6.6**
20	**57.3**	19.9	6.3
40	56.4	19.2	6.0
60	55.6	19.1	5.9

TABLE 7. Ablation study on the threshold of Sentence-BERT similarity. We evaluate on PascalVOC [24], PascalContext [43], and ADE20K [44] and report the mIoU.

threshold	PascalVOC [24]	PascalContext [43]	ADE20K [44]
	20class	59class	150class
-1	56.9	20.2	6.6
0	56.9	20.2	6.6
0.5	79.5	24.4	8.9
0.8	84.8	32.8	15.0

For evaluation, the predicted text T_i is assigned to the ground-truth label that is closest in the Sentence-BERT [46] embedding space. We conduct additional experiments on this assignment metrics. By calculating the IoU only for segments with a cosine similarity above a certain threshold, we enable the evaluation of TAG for different values of similarity. The experimental results are shown in Table 7. The mIoU for segments with thresholds of 0.5 or higher demonstrates performance comparable to supervised, guided open-vocabulary segmentation methods as shown in Table 7. The experimental results also demonstrate that the inferred categories have at least a similarity score of zero or higher within the ground truth label, indicating that the predictions are not entirely off the mark.

V. LIMITATION

While TAG achieves remarkable results, our proposed method still comes with certain limitations. First, as shown in Table 3, TAG depends on the choice of the database, making it challenging to select the optimal database for unknown domains without information on test labels. On the other hand, TAG can flexibly address this limitation by adding new concepts into the database without retraining, unlike language-based methods [14], [16]. Second, TAG does not distinguish between different levels of class granularity. As shown in Figure 6 (e), TAG predicted both 'Porsche' and 'Lamborghini' as 'supercar'. While the predicted categories in the qualitative results are consistently correct, they may not always align with the optimal category desired by the user. Future works might address this issue by considering the frequency of words within the database.

VI. CONCLUSION

In this study, we proposed TAG, Training, Annotation, and Guidance-free open-vocabulary semantic segmentation. TAG employs a novel approach by extracting semantic features from each pixel in an image using CLIP [13], and then retrieving the open-vocabulary categories based on these features from an external database. Through a series of comprehensive experiments and analyses, we have demonstrated the effectiveness and versatility of TAG across various datasets and challenging segmentation tasks. Our results indicate that TAG exhibits robust performance in handling diverse categories, including general classes and fine-grained, proper noun-based segments.

Overall, our findings highlight the potential of TAG as a powerful and effective tool in the field of semantic segmentation. By retrieving the open-vocabulary categories, we have successfully demonstrated the model's capability to handle diverse datasets and open vocabularies without text guidance, paving the way for future advancements and applications in this critical area of computer vision.

REFERENCES

[1] J. Hyun Cho, U. Mall, K. Bala, and B. Hariharan, "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16789–16799.

[2] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[3] H. S. Seong, W. Moon, S. Lee, and J.-P. Heo, "Leveraging hidden positives for unsupervised semantic segmentation," 2023, *arXiv:2303.15014*.

[4] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[5] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.

[7] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7061–7070.

[8] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2955–2966.

[9] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022.

[10] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary universal image segmentation with MaskCLIP," 2022, *arXiv:2208.08984*.

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[12] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[14] P. Rewatbowornwong, N. Chatthee, E. Chuangsuwanich, and S. Suwajanakorn, "Zero-guidance segmentation using zero segment labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1162–1172.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[16] O. Ülger, M. Kulicki, Y. Asano, and M. R. Oswald, "Auto-vocabulary semantic segmentation," 2023, *arXiv:2312.04539*.

[17] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[18] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, N. Peng, L. Wang, Y. Jae Lee, and J. Gao, "Generalized decoding for pixel, image, and language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15116–15127.

[19] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15617–15629.

[20] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3557–3567.

[21] *FreebSD. Web2 Dictionary (Revision 326913)*. Accessed: Feb. 20, 2024. [Online]. Available: https://svnweb.freebsd.org/base/head/share/dict/web2?view=markup&pathrev=326913

[22] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[23] A. Conti, E. Fini, M. Mancini, P. Rota, Y. Wang, and E. Ricci, "Vocabulary-free image classification," 2023, *arXiv:2306.00917*.

[24] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[27] X. Ji, A. Vedaldi, and J. Henriques, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9864–9873.

[28] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7076–7086.

[29] G. Shin, W. Xie, and S. Albanie, "ReCo: Retrieve and co-segment for zero-shot transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 33754–33767.

[30] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[31] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. Jae Lee, "Segment everything everywhere all at once," 2023, *arXiv:2304.06718*.

[32] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero- and few-label semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8248–8257.

[33] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[34] X. Chen, S. Li, S.-N. Lim, A. Torralba, and H. Zhao, "Open-vocabulary panoptic segmentation with embedding modulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1141–1150.

[35] W. Bousselham, F. Petersen, V. Ferrari, and H. Kuehne, "Grounding everything: Emerging localization properties in vision-language transformers," 2023, *arXiv:2312.00878*.

[36] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "ZeroCap: Zero-shot image-to-text generation for visual-semantic arithmetic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17897–17907.

[37] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 2206–2240.

[38] K. Guu, K. Lee, K. Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.

[39] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[40] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen, and A. van den Hengel, "Retrieval augmented classification for long-tail visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6949–6959.

[41] *NLP Library Flair*. Accessed: Feb. 20, 2024. [Online]. Available: https://github.com/flairNLP/flair

[42] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "GroupViT: Semantic segmentation emerges from text supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18113–18123.

[43] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[44] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.

[45] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[46] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

YASUFUMI KAWANO received the master's degree in engineering from Keio University, in 2022, where he is currently pursuing the Ph.D. degree in computer science. His research interest includes machine learning technologies for image and video recognition.

YOSHIMITSU AOKI (Member, IEEE) received the Ph.D. degree in engineering from Waseda University, in 2001. From 2002 to 2008, he was an Associate Professor at the Department of Information Engineering, Shibaura Institute of Technology. He is currently a Professor with the Department of Electronics and Electrical Engineering, Keio University. His research interests include computer vision, pattern recognition, and media understanding.