FOSSIL: Free Open-Vocabulary Semantic Segmentation through Synthetic References Retrieval

Luca Barsellotti[1*] Roberto Amoroso[1*] Lorenzo Baraldi[1] Rita Cucchiara[1,2]
[1]University of Modena and Reggio Emilia, Italy [2]IIT-CNR, Italy
{name.surname}@unimore.it

Abstract

Unsupervised Open-Vocabulary Semantic Segmentation aims to segment an image into regions referring to an arbitrary set of concepts described by text, without relying on dense annotations that are available only for a subset of the categories. Previous works rely on inducing pixel-level alignment in a multi-modal space through contrastive training over vast corpora of image-caption pairs. However, representing a semantic category solely through its textual embedding is insufficient to encompass the wide-ranging variability in the visual appearances of the images associated with that category. In this paper, we propose FOSSIL, a pipeline that enables a self-supervised backbone to perform open-vocabulary segmentation relying only on the visual modality. In particular, we decouple the task into two components: (1) we leverage text-conditioned diffusion models to generate a large collection of visual embeddings, starting from a set of captions. These can be retrieved at inference time to obtain a support set of references for the set of textual concepts. Further, (2) we exploit self-supervised dense features to partition the image into semantically coherent regions. We demonstrate that our approach provides strong performance on different semantic segmentation datasets, without requiring any additional training.

1. Introduction

Semantic segmentation is a classical Computer Vision task that aims at partitioning an image into coherent regions by labeling each pixel according to a predetermined set of categories. The expensive costs of manually annotating training datasets have posed constraints on fully supervised models, restricting their utility to only a narrow set of categories. Consequently, open-vocabulary semantic segmentation has garnered increasing interest in recent years. This paradigm enables models to segment arbitrary categories from free-form textual queries, expanding their applicability to novel real-world scenarios.

The main challenge in open-vocabulary semantic segmentation is how to recognize and localize arbitrary concepts in the image. Previous works focus on learning an alignment in a multi-modal space between textual and dense visual features. A line of research [1, 11, 16, 32–34] exploits dense annotations, that are available for a limited set of categories, to transfer the alignment learned by a large-scale vision-language model (*e.g.*, CLIP [20]) from image to pixel granularity. Instead, other works [4, 30, 31, 37] leverage the vast variability of concepts in large web-crawled sets of image-caption pairs through extensive contrastive learning to bring out the multi-modal alignment in a weakly-supervised fashion. However, a single embedding representing an arbitrary concept is not enough to capture the intra-class variance in the visual appearances of that concept, and such approaches lack explainability.

Few recent works [14, 25] tackle this challenge through the creation of multiple visual prototype embeddings associated with arbitrary concepts, which allow them to avoid the usage of multi-modal feature spaces. Notably, Karazija *et al.* [14] introduces a dynamic strategy where a support set of images is generated on-the-fly. This is achieved by encapsulating textual inputs within a fixed template (*i.e.*, "`A good photo of a` $\langle c_i \rangle$") and providing them to a text-conditioned diffusion model. The resultant support set serves to construct prototype features within the unimodal embedding space of a visual backbone. These prototype features are then employed in the classification of the dense features derived from the input image through a nearest neighbor search in the same embedding space. Nevertheless, generating images at inference time presents notable computational demands. Moreover, the utilization of a pre-determined template constraints the personalization of the textual input, avoiding the formulation of descriptions for arbitrary concepts that encompass their context in the real-world.

To address these problems, we introduce FOSSIL, an architecture that exploits a collection of synthetic visual references that can be retrieved at prediction time to efficiently segment arbitrary textual concepts. Given a large set of web-

*Equal contribution.

crawled captions, we provide them to Stable Diffusion [21] to generate a collection of images. With the recent advances proposed by Tang *et al.* [26], a cross-attention between a word embedding and the dense features within the denoising subnetwork of a diffusion model can be leveraged to extract a heatmap that indicates where that word is active in the generated image. Hence, for each noun in the caption we compute its corresponding heatmap on the generated image and we extract a pair of feature vectors: 1) a Visual Reference Embedding through a visual backbone, pooling its dense features on the most active region in the heatmap, and 2) a Textual Retrieval Embedding, through a text encoder. This embedding is computed as the weighted average between the feature vectors of a pre-determined prompt in which we encapsulate the word and the caption itself. Thus, the Textual Retrieval Embedding mainly represents the noun but considers also the context in which it has been extracted in the caption. At inference time, an arbitrary input concept is embedded using the text encoder to retrieve its most similar Textual Retrieval Embeddings, and, consequently, to obtain their corresponding Visual Reference Embeddings. They can be clustered to produce a set of prototypes in the unimodal space of the visual backbone that represents the arbitrary concept.

Assigning semantic concepts by considering pixel-level features independently would present noisy regions, in particular along borders. However, detecting class-agnostic mask proposals in an unsupervised open-vocabulary setting is a nontrivial problem. To address this challenge, we propose OpenCut, an extension of the recent unsupervised instance segmentation method MaskCut [27]. This method interprets the patches of DINO [3], a self-supervised vision transformer, as nodes of a fully-connected graph and the similarities among their corresponding features as edges of that graph. A fixed threshold is applied to these similarities to compute the optimal cut and bipartition of the graph, creating a mask proposal. In OpenCut, we propose to iteratively shift the threshold to first detect and refine the objects in the foreground and then detect the masks corresponding to the background to cover the majority of the pixels with a unique mask. Thus, we can compute a mask visual feature vector by averaging the dense features that are covered by the same mask, and perform classification at mask-level through prototypes.

Contributions. To sum up, the contributions of this paper are as follows:

- We introduce a novel pipeline for open-vocabulary semantic segmentation, named FOSSIL, that creates a synthetic collection of visual references from a large set of captions using diffusion models and retrieves the reference corresponding to the input text to perform prototype-based segmentation.
- We propose a novel mask proposer approach, named OpenCut, that iteratively bipartitions the features obtained with a self-supervised visual backbone to produce high-quality masks for both foreground and background regions.
- We achieve the new state-of-the-art unsupervised open-vocabulary semantic segmentation performance on 4 segmentation datasets.

2. Related Work

Unsupervised Open-Vocabulary Segmentation. There are two lines of research in the literature on open-vocabulary semantic segmentation. The first exploits a manually annotated training dataset on a fixed set of categories to learn a model capable of generalizing to unseen classes. OpenSeg [11] decouples the task of learning class-agnostic region proposals and aligning multi-modal regions from image-caption pairs. Also, OVSeg [16] proposes a two-stage method, in which a mask proposer produces regions that are given to a fine-tuned CLIP [20] with learnable visual prompts. ODISE [32] directly leverages the internal features of a diffusion model to extract dense features in a multi-modal space. SAN [33] augments a frozen CLIP with a side network that aims to both propose regions and recognize their corresponding textual class.

The second line of research, instead, aims to make the correspondence emerge between visual dense representation and text without the usage of dense annotations. Some works are based on creating an alignment between vision and text in a multi-modal space leveraging a large set of image-caption pairs. GroupViT [30] proposes to hierarchically group semantic regions and align them with the text through contrastive learning. MaskCLIP [37] modifies the CLIP architecture in the last attention layer to align single pixels with text instead of entire images. TCL [4] introduces a grounder on top of a frozen CLIP that learns to ground text to regions through contrastive learning. Other works consider two separate feature spaces and create correspondence between vision and text through external collections. ReCo [25] builds a collection of image representations that can be queried using the input texts, due to the ability of CLIP to compute similarities in the visual space. OVDiff [14] generates on-the-fly a set of visual prototypes through a text-conditioned diffusion model for each input text. Our work positions itself in this research direction and improves the quality of the support set of prototypes by providing both computational and quality gains.

Diffusion models. In recent years, diffusion models have garnered considerable interest due to their capacity to produce high-quality images. In particular, Stable Diffusion [21] represents a lightweight solution that diffuses on a VAE-based latent space instead of raw pixels. In this model, text-conditioning is induced in the denoising subnetwork through

a cross-attention mechanism. In this respect, Tang *et al.* [26] focused on how input words influence the generated image and proposed DAAM, a model that can produce pixel-level heatmaps by upscaling and aggregating cross-attention word-pixel scores. DiffuMask [29] further exploited the advances of DAAM to automatically obtain accurate semantic masks on a synthetic set of images generated by Stable Diffusion. In this paper, we build upon these recent advances to detect objects in a large collection of generated images, extract their dense features, and couple them with text representations to perform open-vocabulary segmentation.

Unsupervised Object Detection and Segmentation. Recently, an area of investigation focused on leveraging the dense features of self-supervised backbones to detect foreground objects in a scene. NCut [24] is an algorithm that interprets the image segmentation task as a fully-connected graph partitioning problem. Each pixel feature is represented as a node and each pair of nodes is connected with an edge that presents a weight equal to the similarity between the corresponding pixel features. To do so, NCut minimizes the energy to partition the graph into two disjoint sets by solving a generalized eigenvalue system defined on the weight matrix and considering the eigenvector corresponding to the second smallest eigenvalue.

On a different note, TokenCut [28] exploits the features of a self-supervised backbone (*i.e.*, the keys from the last attention layer of DINO [3]) as pixel features to apply the NCut algorithm for detecting and segmenting salient objects in images and videos. In particular, a threshold is applied to the weight matrix before computing the eigenvector to enhance the distance between the salient region in the image and the background. CutLER [27] proposes an extension of TokenCut, called MaskCut, to discover multiple objects per image by applying NCut multiple times and masking the weight matrix after each iteration. Thus, the masks extracted on an unlabelled set of images are used to train a supervised segmentation model. We propose to further extend this research direction by iteratively applying MaskCut with different thresholds to segment both foreground and background objects while refining the detected masks.

3. Proposed Method

3.1. Preliminaries

Task definition. Open-Vocabulary Semantic Segmentation aims at segmenting an image $I \in \mathbb{R}^{H \times W \times 3}$ according to a set of arbitrary concepts $c_i \in \mathcal{C}$ described by text. Existing approaches usually tackle the task by associating features extracted from a visual encoder $\Phi_v(I) : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H' \times W' \times D}$ to those extracted from a reference encoder $\Phi_r(c_i) : \mathbb{R}^{d_t} \rightarrow \mathbb{R}^D$, exploiting a similarity function (*e.g.* cosine similarity), so that visual and textual features are treated as lying in a shared multi-modal space. However, a single textual embedding is not sufficient to represent the intra-class variability in the visual appearances of a given concept. Moreover, individually classifying pixel-level features produces noisy semantic regions, especially along borders, whose coherence with the underlying visual object is not guaranteed.

Overview of our approach. To address these weaknesses, we decouple the task into two phases: grouping pixels into visually coherent regions and associating a concept from the set $\mathcal{C}$ to each region. A region proposer in an open-vocabulary setting should be able to detect regions based mostly on visual appearance to maintain a good quality across a large range of concepts. To tackle this challenge, we introduce OpenCut, which aims to iteratively apply MaskCut [27] by varying the threshold τ to accurately detect foreground objects and then the background.

To assign an arbitrary concept to each region, we propose FOSSIL, an architecture that exploits a collection composed of pairs of synthetic Visual Reference Embeddings and Retrieval Textual Embeddings that can be retrieved through an arbitrary textual query and be used to compute similarity against regions in the unimodal visual space. The synthetic visual references are created by providing a large set of captions to Stable Diffusion and, for each noun in the captions, by extracting the corresponding heatmap on the generated image I_g. These heatmaps are binarized to obtain a mask for each noun. The generated image is processed with the visual encoder Φ_v to obtain its dense features and, for each word, a region pooling on the corresponding binary mask is performed to produce a representative Visual Reference Embedding. At the same time, for each noun, a text encoder Φ_t is applied on a pre-determined prompt template in which the noun is inserted and to the caption. Then, the two resulting feature vectors are linearly combined to produce the Retrieval Textual Embedding.

At inference time, the set of textual arbitrary concepts is embedded with the text encoder Φ_t and is used to retrieve the N most similar Textual Retrieval Embeddings. The corresponding Visual Reference Embeddings are clustered to obtain a set of K prototypes. The input image is encoded with the visual encoder Φ_v to obtain its dense features and a set of region proposals is produced using OpenCut. Then, for each region, we perform region pooling on the dense features to create a unique feature vector for that region. Thus, we compute the similarity between the feature vector and the prototypes, to assign the most similar concept to the pixel covered by that region.

3.2. Reference Collection Generation

Our objective is to enable a self-supervised visual backbone to perform open-vocabulary semantic segmentation on an image given a set of free-form arbitrary texts. To achieve so, we want to create a collection of pairs composed of a

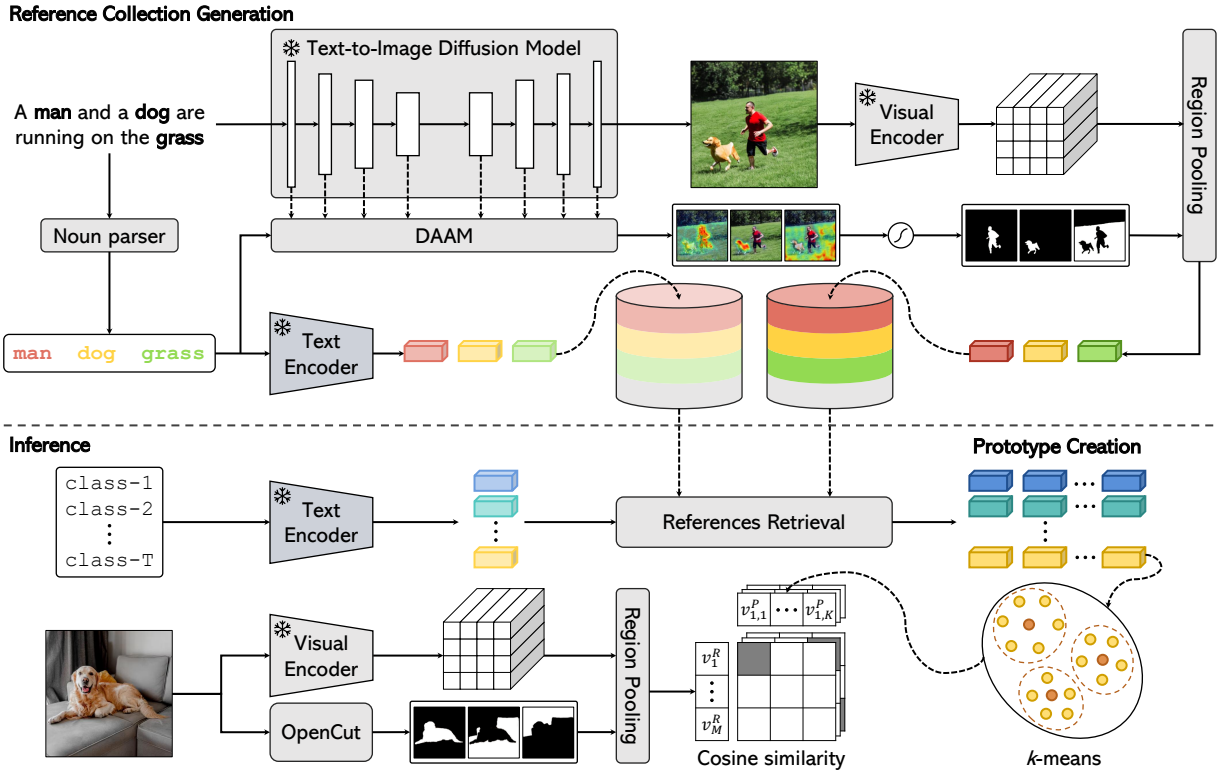

Figure 1. Overview of the proposed FOSSIL approach for training-free Open-Vocabulary Semantic Segmentation.

Visual Reference Embedding, in the backbone space, and a Textual Retrieval Embedding, in the space of a text encoder, for a vast vocabulary from segmentation data. These pairs would couple the visual aspect described by the dense features of the backbone to the corresponding label. However, manually annotated datasets do not cover a vast set of terms and expressions due to the expensive costs of annotating. Hence, we propose to exploit a large web-crawled set of captions that we provide to Stable Diffusion to generate a collection of synthetic images. Thus, we parse nouns from the captions to extract their corresponding heatmap on the generated image through the cross-attention mechanism proposed in DAAM [26]. Since these heatmaps often present peaks on particularly significant portions of the image (*e.g.*, eyes for animals, faces for humans), we apply the sigmoid function on the values of a heatmap followed by a threshold to flatten it. The resulting binary mask can be used to compute the average of the dense features covered by the mask. This produces a Visual Reference Embedding, namely a representative feature vector for the region corresponding to the parsed noun in the caption.

A straightforward approach to building the Textual Retrieval Embedding would be to insert the parsed word in some pre-determined prompt templates, encode them with the text encoder, and compute their average. Nevertheless, this would allow us to only build Textual Retrieval Embeddings for single words, whereas the input text can correspond to any textual description. Hence, we propose to combine the average feature vector of the pre-determined templates with the feature vector of the entire caption. This method moves the vector corresponding to that word towards the real context in which it has been found. Finally, we create an efficient retrieval index on the whole set of collected Textual Retrieval Embeddings.

3.3. Prototype Creation

Given an arbitrary text, we embed the same pre-determined templates used when creating the Textual Retrieval Embeddings using the text encoder to retrieve the N most similar Textual Retrieval Embeddings using the cosine similarity. At inference time, we could interpret the corresponding N Visual Reference Embeddings as prototypes. However, a low value of N risks to be not sufficient to represent the concept, whereas a large N would add outliers that diminish the robustness of the segmentation. So, we cluster the N Visual Reference Embedding through K-Means and we consider the resulting K centroids as prototypes to obtain a trade-off between robustness and representativeness.

3.4. OpenCut

Open-vocabulary segmentation approaches that only leverage pixel-level similarities without considering more

high-level perspectives can produce noisy segmented regions, especially along borders. When multiple objects in a scene come close, indeed, the features along their border embed clues related to multiple semantic elements. On the other hand, implementing region proposal methods in an open-vocabulary setting presents significant threats, as generating high-quality masks for a wide range of concepts requires dense supervision. Following this insight, we propose an extension of the NCut algorithm [24] to provide training-free mask proposals based on the extraction of dense features from a self-supervised backbone.

Preliminaries. Given a dense feature map $\Phi_v(I) \in \mathbb{R}^{H' \times W' \times D}$, NCut builds a fully-connected undirected graph in which each feature vector corresponds to a node, and adds an edge between each pair of features with a weight corresponding to their cosine similarity. A threshold τ is applied on the resulting similarity matrix W to obtain $\bar{W}$ such as

$$\bar{W}_{ij} = \begin{cases} 1e^{-5} & \text{where } W_{ij} < \tau \\ 1 & \text{where } W_{ij} \geq \tau, \end{cases} \tag{1}$$

to binarize and enhance similarities scores. Then, the graph is split into two disjoint graphs by minimizing the energy of the resulting sub-graphs. This operation corresponds to solving the following generalized eigenvalue system

$$(D - \bar{W})x = \lambda x D, \tag{2}$$

and considering the eigenvector x relative to the second smallest eigenvalue λ. In the above equation, D is a diagonal matrix with size $N \times N$ and with $D_{ii} = \sum_j \bar{W}_{ij}$, while $\bar{W}$ is a symmetric matrix with size $N \times N$.

As the resulting eigenvector can be interpreted as a heatmap, we obtain two complementary binary masks by thresholding the eigenvector on its mean value, as

$$M^t_{ij} = \begin{cases} 1 & x_{ij} > \text{mean}(x) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Following MaskCut [27], we first heuristically label the binary mask that contains the patch corresponding to the maximum absolute value in the eigenvector as the foreground mask. Then, in order to obtain the mask relative to the next object, we update the weight matrix by setting to zero the features vector of the nodes corresponding to the current foreground mask and recomputing the weight matrix. This process is repeated until a maximum of t times to detect multiple objects. Also, the procedure is stopped when the thresholded weight matrix is composed of either all 1s or $1e^{-5}$s.

OpenCut. The value of the threshold τ on the weight matrix is determinant in selecting masks with the NCut algorithm, and it is strongly correlated with the structure of the feature space of the visual backbone. Indeed, when considering DINO [3] as backbone, negative or close to 1 values of τ tend to produce masks on background regions rather than foreground objects. In our proposed OpenCut, we leverage this behavior to extract a set of masks so that the majority of pixels in the image are covered. To accomplish this, we iteratively apply the MaskCut algorithm for a set of values of τ, and for each τ we extract a maximum of t masks. The set of chosen τ serves to first identify masks corresponding to foreground objects, refine these masks during iterations, and finally identify the background masks.

Mask refinement. Since each resulting mask is associated with a bipartition of the graph, it is not guaranteed that the mask corresponds to a single region of the image. Hence, we split each binary mask into a mask for each of its connected components. Components that are composed of a number of pixels under a threshold η are discarded to remove noisy regions. During iterations, for each new mask, we check whether it does not present an overlap, measured as Intersection over Union, that is larger than a hyper-parameter μ with a previous mask. If so, it is likely that the two masks correspond to the same object but focus on different parts, and hence we merge them. Moreover, for each new mask, we check whether its surface is not covered for more than a value ρ by the union of the previously extracted masks. If so, we discard the new mask. Then, for each accepted new mask we remove the pixels that are already covered by previously extracted masks. These mechanisms allow us to keep only significant masks and that each pixel is covered by at most a unique mask (*i.e.*, masks are mutually exclusive). There might be uncovered pixels that require to be handled by the segmentation model.

3.5. Inference protocol

Given an image I and a set of arbitrary concepts $\mathcal{C}$ described by texts, we extract the dense features $\Phi_v(I)$ of the image through the visual backbone, and a set of L mutually exclusive binary masks $M_l \in \mathbf{R}^{H \times W}$, $l = 1 \dots L$ using the proposed OpenCut approach. Further, we leverage the procedure described in Section 3.3 to obtain a set of K visual prototypes for each input text. For a binary mask M_l, we upsample it at the resolution $H' \times W'$ of the dense features through bilinear interpolation, and perform a mean-region pooling to construct the region embedding $v_{Rl} \in \mathbf{R}^D$:

$$v_{Rl} = \frac{\sum_{i=1}^{H'} \sum_{j=1}^{W'} M_{lij} \Phi(I)_{ij}}{\sum_{i=1}^{H'} \sum_{j=1}^{W'} M_{lij}}. \tag{4}$$

For a given pixel that is not covered by masks from OpenCut, we consider the patch that covers it as its corresponding mask, with the dense feature vector associated with that patch as its region embedding. For each region embedding v_{Rl}, we compute the cosine similarity against the K prototypes v_{Pk}, $k = 1 \dots K$ of each concept c_a in $\mathcal{C}$, followed

Method	Training Dataset	Support Dataset	Similarity Textual	Similarity Visual	Context	COCOStuff	ADE	Cityscapes
GroupViT [30] [1]	CC12M [5] + RedCaps [9]	-	✓	✗	23.4	15.3	9.2	11.1
MaskCLIP [37] [1]	-	-	✓	✗	26.4	16.4	9.8	12.6
ReCo [25] [1]	-	ImageNet1K [22]	✗	✓	22.3	14.8	11.2	21.1
TCL [4] [1]	CC3M [23] + CC12M [5]	-	✓	✗	30.3	19.6	14.9	23.1
OVDiff [14] [2]	-	-	✗	✓	33.7	-	14.9	-
FOSSIL	-	COCO Captions [6]	✗	✓	**35.8**	**24.8**	**18.8**	**23.2**

Table 1. Comparison with other state-of-the-art unsupervised open-vocabulary semantic segmentation models under the mIoU metric. In "Support Dataset" we report datasets used to create a collection of references. In "Similarity" we report whether the model exploits similarity in a multi-modal embedding space or the unimodal visual space.

by a sigmoid. Finally, we consider the resulting similarity between a region and a concept as the ensembling between the mean of the K similarities against the prototypes and the maximum of them, as follows

$$\bar{s}(v_{Rl}, c_a) = (1-\gamma)\frac{\sum_{k=1}^{K} s(v_{Rl}, v_{Pk})}{K} + \gamma \max_{k=1}^{K} s(v_{Rl}, v_{Pk}), \tag{5}$$

where γ is a weighting hyper-parameter.

4. Experiments

4.1. Implementation Details

For the generation of the Reference Collection, we use all 5 captions per image from COCO Captions [6], Stable Diffusion [21] 2.1 base with 50 diffusion steps and a threshold equal to 0.45 to binarize the heatmaps extracted through DAAM [26]. As visual encoder we use DINOv2 ViT-L/14 [19] on images resized to 512 $\times$ 512, producing dense features $\Phi_v(I) \in \mathbb{R}^{37\times37\times1024}$. As text encoder we use CLIP [20] ViT-L/14 [10] and the set of 7 prompt templates proposed in [20] for zero-shot classification. For building the Textual Retrieval Embedding we adopt a weight equal to 0.9 for the word in the templates and 0.1 for the caption. For MaskCut we use the hyper-parameters proposed in [27]: three stages t on images resized to 480 $\times$ 480 pixels, keys from the last attention layer of a DINO [3] ViT-B/8 backbone as dense self-supervised features, and Conditional Random Field [15] to post-process masks. For mask refinement during the iterations of OpenCut, we use η equal to 16, μ equal to 0.8, and ρ equal to 0.7. We use the faiss library [13] for both efficient retrieval and clustering.

4.2. Results

Datasets. We evaluate the following four benchmarks:

- **Pascal Context** [18] is an extension of the PASCAL-VOC 2010 dataset. It contains 4,996 training images and 5,104 validation images with annotations for 59 classes.
- **COCO Stuff** [2] is an extension of the MS COCO dataset [17] for semantic segmentation. It contains annotations for 171 classes on 118,287 training images and 5,000 validation images.
- **ADE** [35,36] is a challenging segmentation dataset containing indoor and outdoor scenes. It is partitioned into 20,000 training images, 3,000 test images, and 2,000 validation images with annotations for 150 classes.
- **Cityscapes** [8] is a dataset of urban street scenes. It contains 500 validation images with annotations for 19 classes.

Evaluation Protocol. We follow the unsupervised open-vocabulary semantic segmentation evaluation protocol proposed by Cha *et al.* [4]. We use the class names from the default version of MMSegmentation [7] without other modifications. We resize the input image to have a short side of 448 and employ a sliding window approach with a stride of 224 pixels. We use mean Intersection-over-Union (mIoU) to assess the segmentation performance.

Comparison to existing methods. In Table 1 we compare FOSSIL with prior works under the same evaluation protocol: GroupViT [30], MaskCLIP [37], ReCo [25], TCL [4] and OVDiff [14]. As can be seen, our proposal largely outperforms the other methods in all settings, thus confirming the appropriateness of the proposed strategies. Noticeably, our training-free approach overcomes the performances of approaches that employ larger support datasets, and which employ extensive training data.

When comparing across different datasets, we observe a larger margin of improvement on COCOStuff and ADE, respectively of 5.2 and 3.9 mIoU points. Noticeably, these two datasets are the ones with the largest number of classes, respectively 171 and 150, thus underlying the ability of FOSSIL when dealing with a higher number of semantic classes. Further, this also shows that our method is able to

[1] Results from Cha *et al.* [4].
[2] Results from Karazija *et al.* [14].

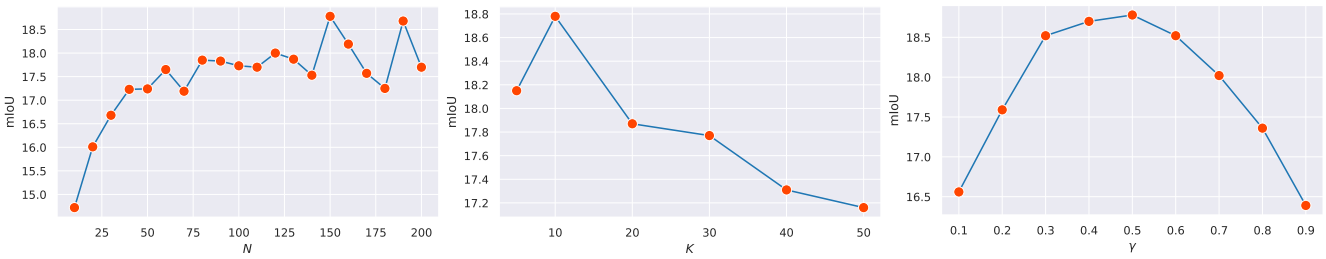

Figure 2. Ablation study on the three inference hyper-parameters N, K and γ on ADE. We test each parameter starting from our best configuration, with $N = 150$, $K = 10$ and $\gamma = 0.5$.

Table 2. Ablation on different visual backbones considering our best configuration on ADE, in terms of mIoU score.

Visual Backbone	Architecture	ADE
MAE [12]	ViT-L/14	2.0
DINO [3]	ViT-B/8	12.0
DINOv2 [19]	ViT-L/14	**18.8**

Table 3. Our best configurations of the hyper-parameters when evaluating each of the 5 benchmarks.

Benchmark	# classes	N	K	γ
Context	59	70	10	0.7
COCOStuff	171	95	25	0.5
ADE	150	150	10	0.5
Cityscapes	19	95	30	0.55

maintain excellent recognition abilities when the number of prototypes in the visual backbone space grows.

Overall, our results show that the contribution provided by the ability to localize concepts generated with Stable Diffusion through DAAM to extract representative dense features largely compensates for the domain shift introduced between synthetic and real images, also due to the high quality reached by diffusion models. Hence, this research direction is proving to be more promising than learning a pixel-level alignment from real images which do not provide locality information for a large vocabulary.

4.3. Ablation Studies

After comparing with other state-of-the-art approaches, we also provide two ablation studies, so as to assess the role of different components of our approach. In particular, we investigate the role of the visual backbone selection and the sensitivity to hyperparameters.

Visual Backbone choice. As the proposed approach is backbone-agnostic, any self-supervised backbone can be employed to build visual prototypes and extract dense features at inference time. To showcase this, and prove the effectiveness of different self-supervised backbones, in Table 2 we report an ablation study on performance obtained on the ADE benchmark with three different backbones: MAE ViT-L/14 [12], DINO ViT-B/8 [3] and DINOv2 ViT-L/14 [19].

As it can be observed, MAE presents a considerably lower performance with respect to the other two backbones, confirming the appropriateness of employing DINO-based backbones. We hypothesize that this is due to the fact that features learned with MAEs have limited semantic coherence when encoding the same concept across different images. When comparing the two considered DINO-based backbones, instead, we notice that DINOv2 largely outperforms DINO, due to a larger architecture (ViT-L/14 instead of ViT-B/8) and to its improved training strategy (as reported in [19]). In the following, we will consider DINOv2 for all the other experiments.

Hyperparameter choice. While being training-free, our approach relies on three main hyper-parameters. Depending on the visual context and the distribution of classes to be detected, these can significantly influence the inference performance and therefore require to be accurately tuned. In particular, these are as follows:

- the number of Textual Retrieval Embedding and Visual Reference Embedding pairs that are retrieved for each arbitrary concept (N);
- the number of visual prototypes obtained as centroids of the K-Means algorithm, applied on the retrieved Visual Reference Embeddings (K);
- the weight attributed to the maximum of the similarities against the set of prototypes of an arbitrary concept, with respect to the mean on that similarities, when computing the concept assigned to a region (γ).

To show that different hyperparameter values can be chosen to obtain a better performance, in Table 3 we report the best configurations obtained on each benchmark. While we did not observe a clear relation between these hyper-parameters and the raw number of classes in a dataset, we hypothesize that their optimal values depend also on other factors, such as the semantic relation in the set of classes and their semantic variance. For instance, in Cityscapes, where classes belong to the urban street domain, a large value of γ may lead

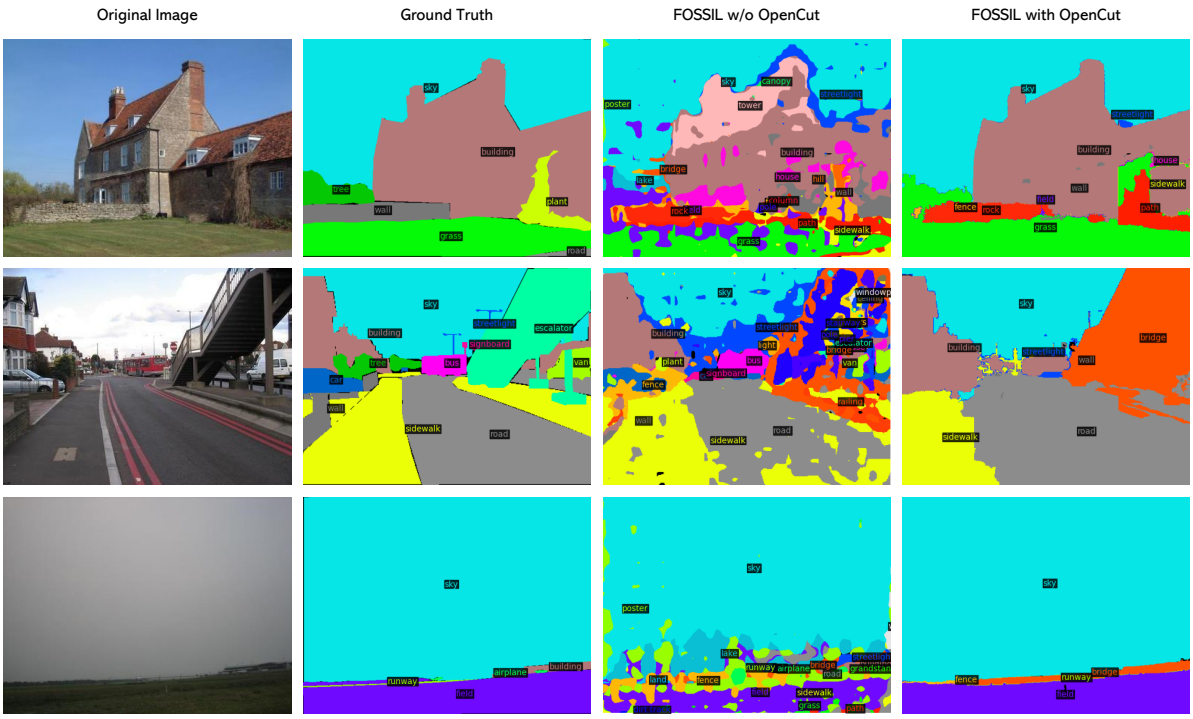

Figure 3. Qualitative results, comparing FOSSIL with and without the OpenCut component.

to assigning outliers introduced during the creation of the Reference Collection, thus resulting in better performance.

Moreover, in Figure 2 we report an ablation study about hyper-parameters N, K, and γ on the ADE benchmark. We vary each parameter one at a time starting from the best configuration. We observe that the parameter that influences performance the most is N, in particular for low values that represent an insufficient amount of references to capture the variance in visual appearances for that class.

4.4. Qualitative results

To complement our evaluation, in Figure 3 we provide a qualitative visualization of the segmentation maps obtained by FOSSIL, on images from the ADE dataset. Here, we also ablate our approach by removing the OpenCut mask proposals and comparing them with our full pipeline, to showcase the role of OpenCut in the final segmentation quality. We firstly observe that FOSSIL is capable of properly segmenting all objects on the scene, assigning them to the correct semantic class, and providing curated segmentation masks that properly align with ground-truth borders. Further, comparing the last two columns of the Figure, the role of the OpenCut component can be clearly observed. As it can be seen, indeed, the mask proposals provided by OpenCut have a high degree of quality, and their adoption results in a cleaner and significantly less noisy result.

5. Conclusion

Open-Vocabulary Segmentation requires an algorithm to segment an input image into regions corresponding to arbitrary textual queries. While previous approaches rely on fine-tuned image-text similarities, in this paper we proposed FOSSIL, an unsupervised open-vocabulary segmentation approach that is training-free and employs collections of visual embeddings to account for visual variety. Our approach leverages text-conditioned diffusion models to generate embeddings that can be efficiently retrieved at prediction time, as a support set to represent textual concepts. Additionally, it employs a self-supervised backbone to partition the image into semantically coherent regions, achieved through an extension of the NCut algorithm that can generate class-agnostic mask proposals. In our experiments, we evaluated our approach on four distinct benchmarks and consistently achieved state-of-the-art results across all of them. Further, we have investigated the sensitivity of hyper-parameters and provided guidelines to choose them. Overall, we demonstrate that open-vocabulary segmentation can be tackled in a training-free manner, by exploiting automatically-generated prototypes which can be retrieved at inference time.

Acknowledgments

This work has been conducted under a research grant co-funded by Leonardo S.p.A. and supported by the PNRR-M4C2 project (PE00000013) “FAIR - Future Artificial Intelligence Research” funded by the European Commission.

References

[1] Luca Barsellotti, Roberto Amoroso, Lorenzo Baraldi, and Rita Cucchiara. Enhancing open-vocabulary semantic segmentation with prototype retrieval. In *Proceedings of the International Conference on Image Analysis and Processing*, 2023. 1

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 6

[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*, 2021. 2, 3, 5, 6, 7

[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*, 2023. 1, 2, 6

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[7] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[8] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 6

[9] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *Advances in Neural Information Processing Systems Datasets and Benchmarks*, 2021. 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 6

[11] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2

[12] 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*, 2022. 7

[13] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 6

[14] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 1, 2, 6

[15] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011. 6

[16] 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*, 2023. 1, 2

[17] 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 *Proceedings of the European Conference on Computer Vision*, 2014. 6

[18] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2014. 6

[19] 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. 6, 7

[20] 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 *Proceedings of the International Conference on Machine Learning*, 2021. 1, 2, 6

[21] 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*, 2022. 2, 6

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 6

[23] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018. 6

[24] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. 3, 5

[25] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 2022. 1, 2, 6

[26] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. 2, 3, 4, 6

[27] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 5, 6

[28] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[29] 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. 3

[30] 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*, 2022. 1, 2, 6

[31] 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*, 2023. 1

[32] 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*, 2023. 1, 2

[33] 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*, 2023. 1, 2

[34] 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 *Proceedings of the European Conference on Computer Vision*, 2022. 1

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 6

[36] 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*, 2019. 6

[37] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2, 6

FOSSIL: Free Open-Vocabulary Semantic Segmentation through Synthetic References Retrieval Supplementary Material

In this supplementary material we:

- provide additional implementation details on the proposed method;
- provide a qualitative evaluation of the Reference Collection Generation step, focusing on the heatmaps and the binary masks extracted for each generated concept;
- extend our qualitative evaluation of predictions made on the benchmarks considered, providing a more comprehensive view of the performance of our model.

A. Additional Implementation Details

A.1. Efficient Retrieval and Clustering

To optimize the execution of retrieval and clustering procedures, we harness the capabilities of the `faiss` library [3]. Specifically, our approach leverages a retrieval index based on the Hierarchical Navigable Small World graph exploration (HNSW) [4] technique. This method employs an approximation of the nearest neighbor search, thereby enhancing the efficiency of retrieving Textual Retrieval Embeddings and Visual Reference Embeddings.

A.2. Textual Prompt Templates

During both the Reference Collection Generation and Prototype Creation phases, we encapsulate arbitrary textual concepts using pre-defined prompt templates. These templates are the ones introduced in CLIP [6], as follows:

- `itap of a {}.`
- `a bad photo of the {}.`
- `a origami {}.`
- `a photo of the large {}.`
- `a {} in a video game.`
- `art of the {}.`
- `a photo of the small {}.`

B. More Qualitatives

B.1. Reference Collection Generation

In Figure 1, we present a visual representation of the qualitative results obtained during the Reference Collection Generation step. In particular, we illustrate: 1) the caption used to condition Stable Diffusion [7], 2) the resulting generated image, 3) the heatmaps corresponding to nouns, extracted through DAAM [8], and 4) the binarized heatmaps used to perform region pooling on the dense features of the visual backbone. These qualitatives show the effectiveness of DAAM in localizing words in the generated image. The resulting heatmaps can be thresholded to produce approximate binary masks for the identified concept. It is worth noting that minor inaccuracies in border delineation do not significantly impact the resulting feature vector, as these masks are employed for feature averaging.

B.2. Prediction Qualitatives

In Figure 2, we present qualitative results depicting the predictions made by our method on three benchmark datasets: PASCAL Context [5], Cityscapes [2] and COCOStuff [1]. Additionally, we provide a comparison by displaying the same predictions without the utilization of OpenCut, emphasizing the impact of OpenCut in refining masks. Specifically, the showcased qualitative results underscore the robust recognition capabilities of FOSSIL in identifying semantic elements within the scenes. However, it is noteworthy that the integration of OpenCut is essential for refining the resulting segments, particularly in areas adjacent to borders.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3

[2] 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*

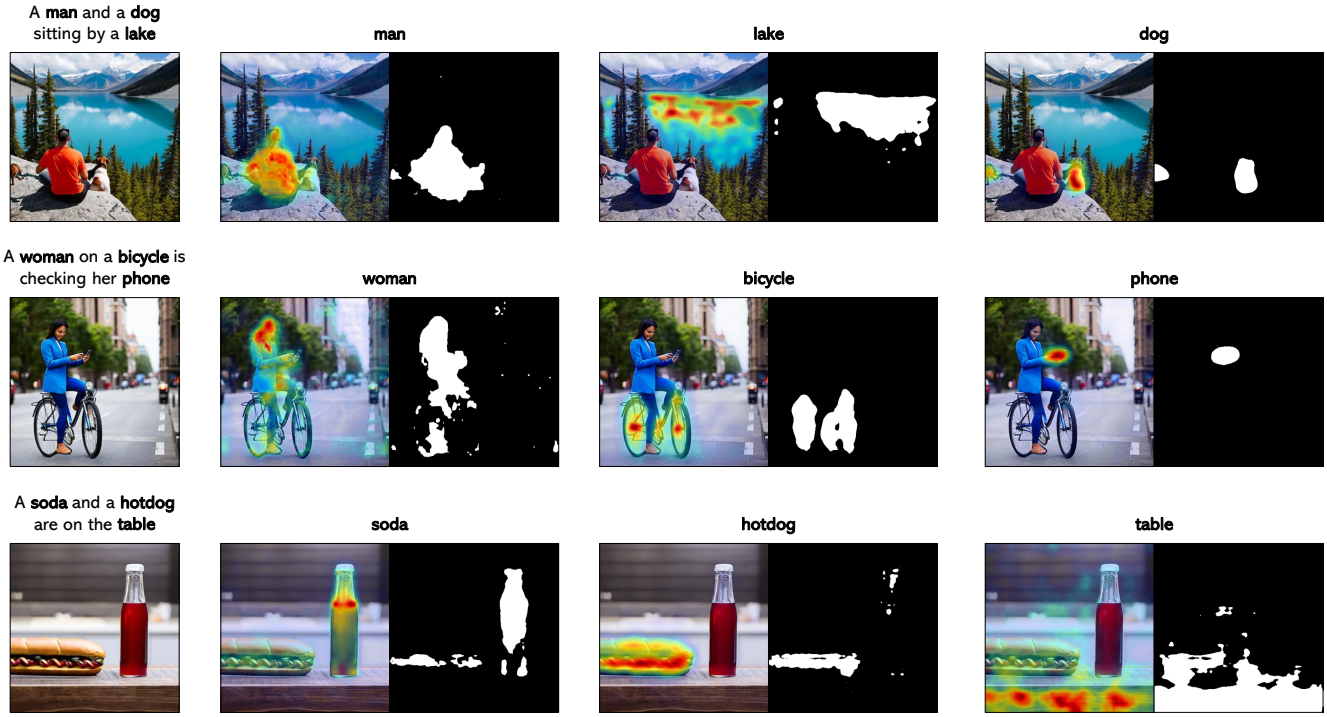

Figure 1. Qualitative results of the Reference Collection step. On the left, we report three captions used to condition Stable Diffusion and generate the corresponding images. On the right, we report the heatmap obtained through DAAM [8] for each noun from the caption and the corresponding binary mask.

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016. 1, 3

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 1

[4] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1

[5] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2014. 1, 3

[6] 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 *Proceedings of the International Conference on Machine Learning*, 2021. 1

[7] 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*, 2022. 1

[8] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. 1, 2

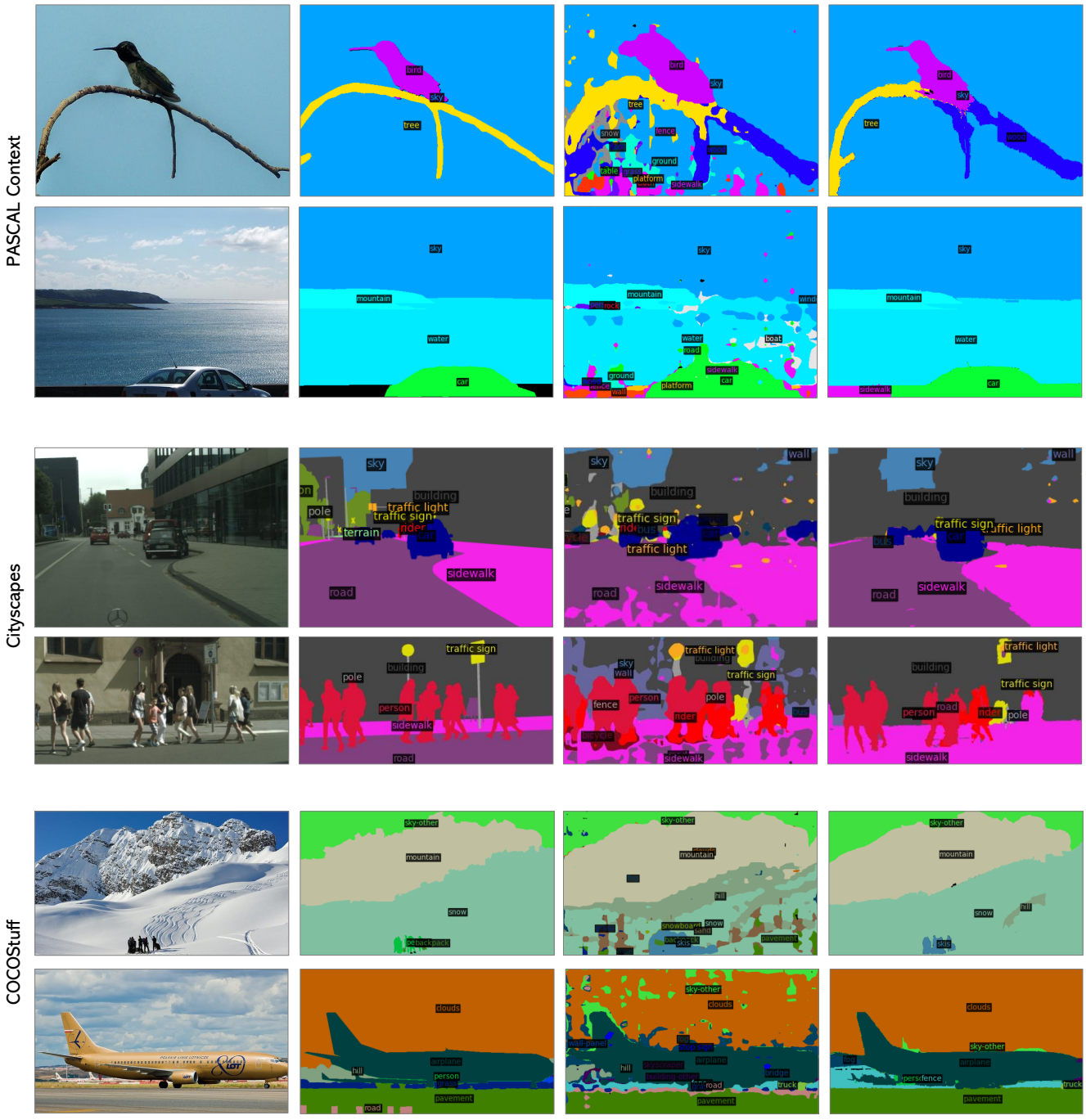

Figure 2. Qualitative results, comparing FOSSIL with and without the OpenCut component on the PASCAL Context [5], Cityscapes [2] and COCOStuff [1].