

Delving into Shape-aware Zero-shot Semantic Segmentation

Xinyu Liu^{1,2}, Beiwen Tian^{2,4}, Zhen Wang³, Rui Wang³, Kehua Sheng³, Bo Zhang³,
Hao Zhao², Guyue Zhou²

¹Xidian University ³Didi Chuxing

²Institute for AI Industry Research (AIR), Tsinghua University

⁴Department of Computer Science and Technology, Tsinghua University

liuxinyu@stu.xidian.edu.cn, zhaohao@air.tsinghua.edu.cn

Abstract

Thanks to the impressive progress of large-scale vision-language pretraining, recent recognition models can classify arbitrary objects in a zero-shot and open-set manner, with a surprisingly high accuracy. However, translating this success to semantic segmentation is not trivial, because this dense prediction task requires not only accurate semantic understanding but also fine shape delineation and existing vision-language models are trained with image-level language descriptions. To bridge this gap, we pursue **shape-aware** zero-shot semantic segmentation in this study. Inspired by classical spectral methods in the image segmentation literature, we propose to leverage the eigen vectors of Laplacian matrices constructed with self-supervised pixel-wise features to promote shape-awareness. Despite that this simple and effective technique does not make use of the masks of seen classes at all, we demonstrate that it outperforms a state-of-the-art shape-aware formulation that aligns ground truth and predicted edges during training. We also delve into the performance gains achieved on different datasets using different backbones and draw several interesting and conclusive observations: the benefits of promoting shape-awareness highly relates to mask compactness and language embedding locality. Finally, our method sets new state-of-the-art performance for zero-shot semantic segmentation on both Pascal and COCO, with significant margins. Code and models will be accessed at [SAZS](#).

1. Introduction

Semantic segmentation has been an established research area for some time now, which aims to predict the categories of an input image in a pixel-wise manner. In real-world applications including autonomous driving [18], medical diagnosis [32, 47] and robot vision and navigation [9, 64], an accurate semantic segmentation module provides a pixel-wise

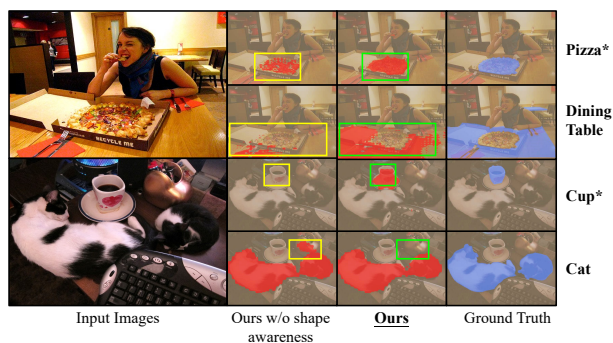


Figure 1. Without retraining, SAZS is able to precisely segments both seen and unseen objects in the zero-shot setting, largely outperforming a strong baseline. * denotes unseen categories during training).

understanding of the input image and is crucial for subsequent tasks (like decision making or treatment selection).

Despite that significant progress has been made in the field of semantic segmentation [6, 7, 33, 50, 53, 55, 58, 60, 62], most existing methods focus on the closed-set setting in which dense prediction is performed on the same set of categories in training and testing time. Thus, methods that are trained and perform well in the closed-set setting may fail when applied to the open world, as pixels of unseen objects in the open world are likely to be assigned categories that are seen during training, causing catastrophic consequences in safety-critical applications such as autonomous driving [63]. Straightforward solutions include fine-tuning or retraining the existing neural networks, but it is impractical to enumerate unlimited unseen categories during retraining, let alone large quantities of time and efforts needed.

More recent works [4, 15, 25, 28, 41] address this issue by shifting to the zero-shot setting, in which the methods are evaluated with semantic categories that are unseen during training. While large-scale pre-trained visual-language models such as CLIP [42] or ALIGN [19] shed light on the

potential of solving zero-shot tasks with priors contained in large-scale pre-trained model, how to perform dense prediction task in this setting is still under-explored. One recent approach by Li et.al. [25] closes the gap by leveraging the shared embedding space for languages and images, but fails to effectively segment regions with fine shape delineation. If the segmented shape of the target object is not accurate, it will be a big safety hazard in practical applications, such as in autonomous driving.

Inspired by the classical spectral methods and their intrinsic capability of enhancing shapeawareness, we propose a novel **Shape-Aware Zero-Shot** semantic segmentation framework (**SAZS**) to address the task of zero-shot semantic segmentation. Firstly, the framework enforces vision-language alignment on the training set using known categories, which exploits rich language priors in the large-scale pre-trained vision-language model CLIP [42]. Meanwhile, the framework also jointly enforces the boundary of predicted semantic regions to be aligned with that of the ground truth regions.

Lastly, we leverage the eigenvectors of Laplacian of affinity matrices that is constructed by features learned in a self-supervised manner, to decompose inputs into eigensegments. They are then fused with learning-based predictions from the trained model. The fusion outputs are taken as the final predictions of the framework.

As illustrated in Fig. 1, compared with [25], the predictions of our approach are better aligned with the shapes of objects.

We also demonstrate the effectiveness of our approach with elaborate experiments on PASCAL-5ⁱ and COCO-20ⁱ, the results of which show that our method outperforms former state-of-the-arts [4, 25, 37, 38, 52, 54] by large margins. By examining a) the correlation between **shape compactness** of target object and IoU and b) the correlation between the **language embedding locality** and IoU, we discover the large impacts on the performance brought by the distribution of language anchors and object shapes. Via extensive analyses, we demonstrate the effectiveness and generalization of SAZS framework’s shape perception for segmenting semantic categories in the open world.

2. Related works

2.1. Zero-Shot Semantic Segmentation

The main goal of the zero-shot semantic segmentation task(ZSS) is to perform pixel-wise predictions for objects that are unseen during training. Recent works on ZSS have seen two main branches: the generative methods and the discriminative methods.

The generative methods [4, 15, 28] produce synthesized features for unseen categories.

ZS3Net [4] utilizes a generative model to create a visual

representation of objects that were not present in the training data. This is achieved by leveraging pre-trained word embeddings. CaGNet [15] highlights the impact of contextual information on pixel-level features through a network learning to generate specific contextual pixel-level features. In CSRI [28], constraints are introduced to the generation of unseen visual features by exploiting the structural relationships between seen and unseen categories.

As for the discriminative methods, SPNet [54] leverages similarities between known categories to transfer learned representations to other unknown categories. Baek et al. [2] employ visual and semantic encoders to learn a joint embedding space with the semantic encoder converting the semantic features into semantic prototypes. Naoki et al. [22] introduce variational mapping by projecting the class embeddings from the semantic to the visual space. Lv et al. [34] present a transductive approach using target images to mitigate the prediction bias towards seen categories. LSeg [25] proposes a language-driven ZSS model, mapping pixels and names of labels into a shared embedding space for pixel-wise dense prediction.

Though much pioneering efforts have been spent, the dense prediction task requires fine shape delineation while most existing vision-language models are trained with image-level language descriptions. How to effectively address these problems is the focus of our work.

2.2. Shape-aware Segmentation

Shape awareness is beneficial to dense prediction tasks. Most of the semantic segmentation methods [6, 33, 45, 53] cannot preserve object shapes since they only focus on feature discriminativeness but ignore proximity between central and other positions.

Meanwhile, SGSNet [59] takes a hierarchical approach to aggregating the global context when modeling long-range dependencies, considering feature similarity and proximity to preserve object shapes. ShapeMask [23] refines the coarse shapes into instance-level masks. The shape priors provide powerful clues for prediction. Gated-SCNN [49] proposes a two-stream architecture for semantic segmentation that explicitly captures shape information as a separate processing branch. The key point is to enable the interactive flow of information between the two networks, allowing the shape stream to focus on learning and processing of edge information. Liu et al. [31] construct the spatial propagation networks for learning the affinity matrix. The affinity matrix allows a tractable modeling of the dense, global pairwise relationships of pixels.

2.3. Spectral Methods for Segmentation

Among different segmentation schemes, spectral methods for clustering employ the eigenvectors of a matrix derived from the distance between points, which have been

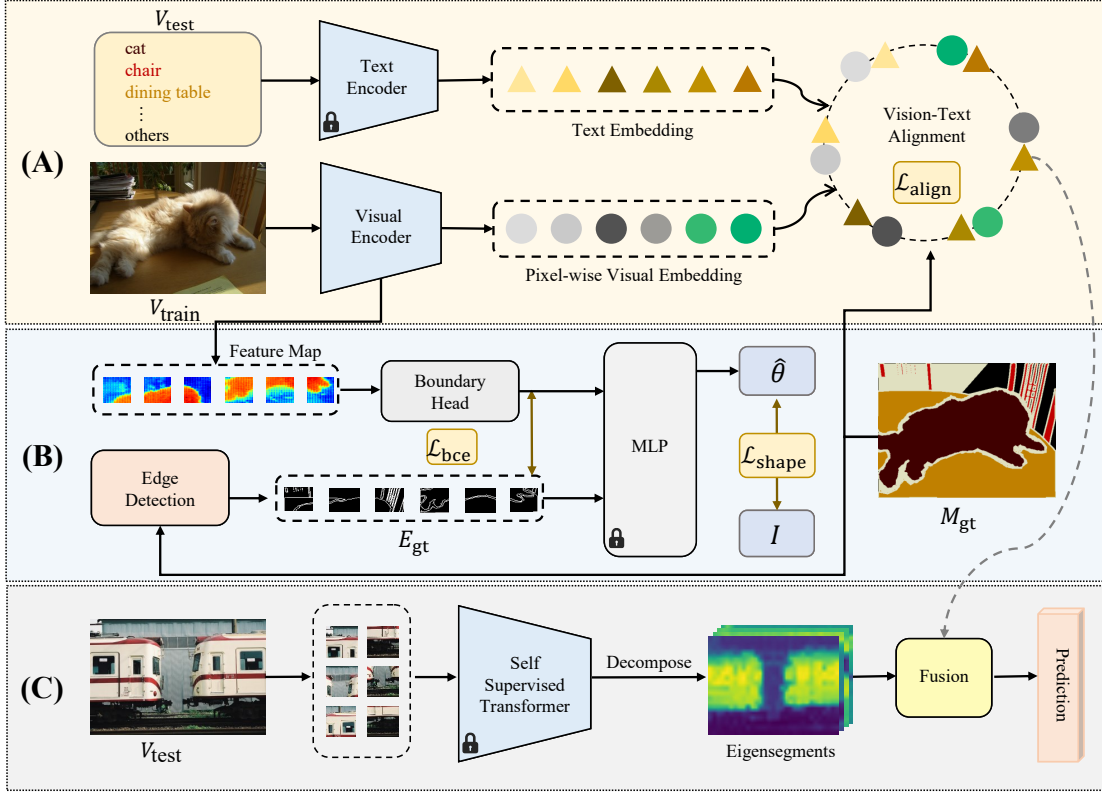


Figure 2. The overview of SAZS framework. SAZS addresses the task of zero-shot semantic segmentation, which aims to segment the test set image V_{test} by open-set categories without additional training of the network. During training, (A) the input image V_{train} is transformed into pixel-wise visual embeddings which are aligned with the text embeddings of training categories T_{train} , according to the ground-truth semantic maps M_{gt} . The text embeddings are obtained by the pre-trained text encoder of CLIP [42] and serve as optimization anchors of the CLIP feature space. (B) In order to aggregate shape priors contained in the input image, SAZS jointly trains on the constraint task of boundary detection by comparing the ground-truth boundaries and the predictions of boundary heads of the visual encoder. (C) During inference, in order to reduce the domain gap between seen and unseen categories, SAZS fuses the pixel-wise predictions of the neural networks with eigensegments obtained by non-learning-based spectral analysis. Note that, modules marked with the lock icon are pre-trained and not optimized during training of SAZS.

successfully used in many applications.

Shi et al. [48] regard image segmentation as a graph partitioning problem, which proposes a novel global criterion, the normalized cut, to segment the image. Soft segmentations [1] are generated automatically by fusing high-level and low-level image features in a graph structure. The purpose of constructing this graph is to enable the corresponding Laplacian matrix and its eigenvectors to reveal semantic objects and soft transitions between objects. In our work, we utilize the eigensegments obtained by self-supervised spectral decomposition with the network outputs as the framework’s predictions to avoid the bias of the learning-based model on the training set and further improve shape-awareness.

2.4. Vision-Language Modeling

An extensive group of works have investigated the zero-shot semantic segmentation task. The key idea behind many of these works is to exploit priors encoded in pre-trained word embeddings to generalize to unseen classes and achieve dense predictions [4, 13, 15, 17, 20, 22, 25, 27–29, 39, 54, 61], such as word2vec [36], GloVe [40] or BERT [10]. CLIP [42] has recently demonstrated impressive zero-shot generalization in various image-level classification tasks. As a result, several works have since exploited the vision-language embedding space learned by CLIP [42] to enhance dense prediction capabilities [25, 44, 56]. CLIP develops contrastive learning with a large-capacity language and visual feature encoder to train extremely robust models for zero-shot image classification. But the performance of large-scale pre-trained vision encoders transferred

to pixel-level classification work is unsatisfactory. Unfortunately, the direct utilization of the extracted image-level vision-language features ignores the discrepancy between image-level and the pixel-level dense prediction task, the latter of which is the focus of our work. According to our study, the shape-aware prior and supervision can bridge this discrepancy and get more accurate segmentation results.

3. Methods

The goal of zero-shot semantic segmentation is to extend semantic segmentation task to the unseen categories other than those in the training datasets. One potential approach to introduce extra priors is to leverage pre-trained vision-language models, yet most of these models focus on the image-level prediction and fail to transfer to dense prediction tasks.

To this end, we propose a novel method named **Shape-Aware Zero-Shot Semantic Segmentation (SAZS)**.

This approach leverages the rich language priors contained in the pre-trained CLIP [42] model, while also exploiting the proximity between local regions to perform the boundary detection task with constraints. Meanwhile, we utilize spectral decomposition of self-supervised visual features to improve our approach’s sensitivity to shape, and integrate this with pixel-wise prediction.

The overall pipeline of our methods is depicted in Fig. 2. The input image is first transformed by an image encoder into pixel-wise embeddings, which are then aligned with precomputed text embeddings obtained by the text encoder of pre-trained CLIP model (Part A in Fig. 2). Meanwhile, an extra head in the image encoder is used to predict the boundaries in patches, which are optimized towards the ground-truth edges obtained from segmentation ground truths (Part B in Fig. 2). In addition, we further exploit proximity of local regions during inference by decomposing the image by spectral analysis and fusing the output eigensegments with class-agnostic segmentation results (Part C in Fig. 2).

In the following section, we first formally define the addressed task and introduce the notations in Sec. 3.1. Then we describe the loss designs for the vision-language alignment and boundary prediction in Sec. 3.2 and Sec. 3.3, respectively. The inference pipeline involving spectral decomposition of the proposed affinity matrix is introduced in Sec. 3.4.

3.1. Task Definition

Following HSNet [37], we denote the training set by $\mathcal{D}_{\text{train}} = \{(I, M, \mathcal{S})\}$ and testing set by $\mathcal{D}_{\text{test}} = \{(I, M, \mathcal{U})\}$, where $I \in \mathbb{R}^{H \times W \times 3}$ and $M \in \mathbb{R}^{H \times W \times C}$ denote an input image and the corresponding ground-truth semantic mask with digit encoding. \mathcal{S} denotes the set of K

potential labels in I , while \mathcal{U} denotes the set of unseen categories during testing. The two sets are strictly exclusive in our setting (i.e., $\mathcal{S} \cap \mathcal{U} = \emptyset$).

Before inferencing on $\mathcal{D}_{\text{test}}$ targeting \mathcal{U} , the model is trained on $\mathcal{D}_{\text{train}}$ with ground-truth labels from \mathcal{S} . This means the categories in the test set are never seen during training, making the task formulated in a zero-shot setting. Once the model is well-trained, it is expected to generalize to unseen categories and achieve high performance for dense prediction of target objects in the open world.

3.2. Pixel-wise Vision-Language Alignment

Comparing distances between pixel features and different text anchor features in the shared feature space is a straightforward approach for zero-shot semantic segmentation. However, while the pioneer work CLIP [42] introduces a shared feature space for visual and text inputs, the image-level CLIP visual encoder is infeasible for dense prediction tasks since fine details in images, as well as the correlation between pixels, are lost. In this section, we describe our approach to address this issue by optimizing a dense visual encoder separate of CLIP and enforcing the pixel-wise output features towards the text anchors in the CLIP feature space during training.

Visual Encoder We employ Dilated residual networks (DRN) [57] and Dense Prediction Transformers (DPT) [43] to encode images into pixel-level embeddings. More specifically, an input image of size $H \times W \times 3$ is first processed with standard augmentation to $\tilde{H} \times \tilde{W} \times 3$ and then passed as input to the visual encoder, resulting in a feature map $\mathcal{F}_V \in \mathbb{R}^{\tilde{H} \times \tilde{W} \times D}$, where D is the feature size in DPT.

Text Encoder While most concurrent methods use digit labels (e.g., 0, 1, 2) to represent categories, we take embeddings of the category names (e.g. "airplane", "cat") as the anchors of feature space. These embeddings are obtained with the CLIP text encoder. Specifically, we adopt the pre-trained CLIP text encoder to map the names of K categories from \mathcal{S} into CLIP feature space as the anchor features $\mathcal{F}_T \in \mathbb{R}^{K \times D}$, which is later used as targets for optimization.

Note that, the visual features \mathcal{F}_V and the text features \mathcal{F}_T have the same dimension D .

Vision-Language Alignment To enforce vision-language alignment, the distances between pixels and corresponding semantic category should be minimized while the distances between pixels and other categories should be maximized. Under the assumption that pixel-wise vision and language features are embedded in the same feature space, we leverage the cosine similarity $\langle \cdot, \cdot \rangle$

as the quantitative distance metric between features and propose the alignment loss as the sum of cross entropy losses over seen classes of all pixels:

$$\mathcal{L}_{\text{align}} = \sum_{i,j}^{\tilde{H}, \tilde{W}} \left(-\log \frac{e^{\langle \mathcal{F}_V[i,j], \mathcal{F}_T[k_{ij}] \rangle}}{\sum_{k'=1}^{|S|} e^{\langle \mathcal{F}_V[i,j], \mathcal{F}_T[k'] \rangle}} \right) \quad (1)$$

In Eq. 1, $\mathcal{F}_V[i,j]$ denotes pixel visual feature at position (i,j) , $\mathcal{F}_T[k]$ denotes k -th text anchor features and k_{ij} denotes index of ground-truth category of pixel at (i,j) .

3.3. Shape Constraint

Since CLIP is trained on an image-level task, simply leveraging the priors in the CLIP feature space may be insufficient for dense prediction tasks. To address this issue, we introduce boundary detection as a constraint task, so that the visual encoder is able to aggregate finer information contained in images. Inspired by InverseForm [3], we address this constraint task by optimizing the affine transformation between ground-truth edges and edges in feature maps towards identity transformation matrix.

More specifically, as shown in Fig. 2, we extract middle-layer features of the visual encoders and split them into patches. On the one hand, the ground truth edges within the patches are obtained by applying Sobel operator on ground truth semantic masks. On the other hand, the feature patches are processed by a boundary head. Then, we calculate the affine transform matrix $\hat{\theta}_i$ for the i -th patch between ground-truth edges and processed feature patches with a pre-trained MLP. Note that, this MLP is trained in advance with edge masks and not optimized during our method’s training. We optimize this affine transform matrix towards identity matrix by:

$$\mathcal{L}_{\text{shape}} = \frac{1}{T} \sum_{i=1}^T \left| \hat{\theta}_i - I \right|_F \quad (2)$$

where T denotes the number of patches and $|\cdot|$ denotes Frobenius norm.

Furthermore, we directly calculate the binary cross entropy loss \mathcal{L}_{bce} between the predicted edge masks of the whole image and corresponding ground truths to further optimize the performance of boundary detection.

After jointly training on the task of boundary detection, the visual encoder is enabled to collect and leverage shape priors in the input images. Ablation studies detailed later show that shape awareness introduced by $\mathcal{L}_{\text{shape}}$ and \mathcal{L}_{bce} brings about notable improvements.

Finally, the overall loss to optimize during training is:

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_1 \mathcal{L}_{\text{shape}} + \lambda_2 \mathcal{L}_{\text{bce}} \quad (3)$$

where λ_1 and λ_2 are loss weights.

3.4. Self-supervised Spectral Decomposition

We seek to decompose the input images into eigensegments with clear boundaries in an unsupervised manner, and then fuse these eigensegments with the predictions of the neural networks in the fusion module in Fig. 2.

The derivation of affinity matrix is the key to spectral decomposition. Following Melas-Kyriazi et al. [35], we first leverage the features f from the attention block of the last layer of a pre-trained self-supervised transformer (i.e., DINO [5]). The affinity between pixel i and j is defined as:

$$Z_{\text{sem}}(i,j) = f_i \cdot f_j^T \quad (4)$$

Note that, the self-supervised transformer is only used during inference and its weights are not optimized.

While the affinities derived from transformer features are rich in semantic information, the low-level proximity including color similarity and spatial distance is missing. Inspired by image matting [8, 24], we first transform the input image into the HSV color space: $X(i) = (\cos(h), \sin(h), s, v, x, y)_i$, where h, s, v are the respective HSV coordinates and (x, y) are the spatial coordinates of pixel i . Then, the affinity between pixels is defined as

$$Z_{\text{shape}}(i,j) = 1 - \|X(i) - X(j)\|_2, \quad j \in \text{KNN}(i) \quad (5)$$

where $\|\cdot\|_2$ denotes 2-norm. The overall affinity matrix is defined as the weighted sum of the two:

$$Z(i,j) = Z_{\text{sem}} + \lambda \cdot Z_{\text{shape}} \quad (6)$$

With the affinity matrix, we now can compute the eigenvectors of the Laplacian L of the affinity matrix, which are used to decompose the image into multiple eigensegments.

3.5. Inference

Given an image for inference, we first encode the phrases of the categories using the pre-trained text encoder CLIP and obtain textual features $\mathcal{F}_T \in \mathbb{R}^{C \times D}$ for C categories, each of which is represented by a D -dimension embedding. Then we leverage the trained visual encoder to obtain the visual feature map $\mathcal{F}_V \in \mathbb{R}^{\tilde{H} \times \tilde{W} \times D}$. The final logits $\hat{F}_{ij} = \mathcal{F}_V(i,j) \cdot \mathcal{F}_T^T$ are calculated as the cosine similarities between the visual feature map and textual features. In the mean time, we employ the pre-trained DINO to extract semantic features in an unsupervised manner and calculate the top K spectral eigensegments E_k ($K = 5$ in our implementation). The final prediction results are generated by the fusion module, which selects from the sets of predictions according to the maximal IoU (denoted as Φ_{FUSE}) of the E_k and $\text{argmax } \hat{F}_{ij}$.

$$\text{Pred}_{ij} = \Phi_{\text{FUSE}} \left(E_k, \text{argmax } \hat{F}_{ij} \right) \quad k \in \{0, 1, \dots, K\} \quad (7)$$

4. Experiments

4.1. Datasets

We extensively evaluate our method on two datasets dedicated for the task of zero-shot semantic segmentation: PASCAL-5ⁱ [12] and COCO-20ⁱ [30]. Built upon PASCAL VOC 2012 [12] and augmented by SBD [16], PASCAL-5ⁱ contains 20 categories which are further divided into 4 folds denoted by 5⁰, 5¹, 5² and 5³. Each image is annotated with 5 categories within each fold. Similarly, based on MS COCO [30], COCO-20ⁱ is a more challenging dataset with 80 categories divided into four folds denoted by 20⁰, 20¹, 20² and 20³, and each of the four folds contains 20 categories. Of the four folds in the two datasets, one is used for evaluation (i.e., the target fold) while the other three are used for training. In the following section, we denote each experiment setup by the target fold.

Following prior literature on zero-shot semantic segmentation, we adopt mean intersection over union (mIoU) and foreground-background IoU (FBIoU) as the evaluation metrics. Specifically, mIoU is the average of IoUs of the categories in target fold and FBIoU is the average of foreground IoU and background IoU.

4.2. Implementation Details

In our experiments, we employ the pre-trained CLIP-ViT-B/32 as the text encoder. Background or unknown category is regarded as "others" when mapped from text to CLIP features. The visual encoder is implemented by DRN [57] or DPT [43] with ViT [11] as the backbone. When training on the task of boundary detection, each feature map for the shape boundary and the corresponding ground truth are splitted into 3×6 patches. Each patch pair is then fed into the MLP in Part B of Fig. 2 to calculate the affine transformation matrix.

During training, the network is optimized by an SGD optimizer with a momentum of 0.95 and a learning rate of 5×10^{-5} decayed by a polynomial scheduler. With ViT as the backbone of visual encoder, the training process finishes within 5 epochs on 4 NVIDIA Tesla V100 GPUs with a batch size of 6.

4.3. Results

The proposed method SAZS has been evaluated on the PASCAL-5ⁱ and COCO-20ⁱ datasets under zero-shot settings, alongside several baselines for comparison. The performances are reported in Tab. 1. With DRN as the visual encoder backbone, our method achieves large margins over the strong baseline LSeg [25], with mIoU improved by 6.1% and 4.8% on PASCAL-5ⁱ and COCO-20ⁱ respectively. Our model also outperforms LSeg [25] by large margins with the ViT backbone underlying DPT, with mIoU improved by 7.2% and 11.2%. The performance enhance-

ments of SAZS remain consistent across different visual encoder choices, highlighting its effectiveness.

In addition, we conduct cross-dataset validation by training on the COCO-20ⁱ dataset and testing on PASCAL-5ⁱ. As shown in Table 2, our method outperforms OpenSeg [14] and LSeg+ [25] in zero-shot dense prediction tasks with clear margins. It is worth noting that all three methods are trained on a larger semantic segmentation dataset (COCO-20ⁱ). These performance gaps demonstrate the generalization ability of our shape-aware training framework across datasets. We also provide qualitative results for the proposed method SAZS. in Fig. 3 and Fig. 4. In these figures, we illustrate the predictions of SAZS with and without shape awareness on COCO-20ⁱ and PASCAL-5ⁱ respectively, showing its ability to make precise predictions on both seen and unseen categories.

4.4. Ablation Study

To further demonstrate the effectiveness of design choices in our approach, we perform detailed ablation studies by evaluating our method with or without shape constraint during training as well as the fusion of network predictions with eigensegments. Results on PASCAL-5ⁱ are reported in Tab. 3 and results on COCO-20ⁱ are reported in Tab. 4 and Tab. 5.

Effects of Shape-awareness The motivation for auxiliary constraint $\mathcal{L}_{\text{shape}}$ is to learn the shape priors of images contained in the target boundaries. We observe that without training on the constraint task of boundary detection, the performances of the proposed method tend to decline. Specifically as reported in Tab. 4 and Tab. 5, the mIoU of SAZS drops by 1.4% and by 1.5% with ViT and DRN backbone on COCO-20ⁱ when training without $\mathcal{L}_{\text{shape}}$. The performance gaps clearly indicate the significant role played by shape-awareness in the proposed SAZS framework.

Effect of Fusion with Spectral Eigensegments We also demonstrate the importance of fusing with spectral eigensegments during inference. Without the fusion module, the mIoU dramatically decreases by 7.0% on PASCAL-5ⁱ and by 6.2% (ViT backbone) and 8.6% (DRN backbone) on COCO-20ⁱ, as reported in Tab. 3, Tab. 4 and Tab. 5. These large margins indicate that eigensegments obtained by spectral decomposition of the affinity matrices largely suppress the bias on the training dataset and seen categories.

4.5. Ablation of Z_{sem} and Z_{shape}

We conduct an ablation experiment on the PASCAL-5ⁱ dataset to investigate the effects of Z_{sem} and Z_{shape} in our fusion module. As shown in Table 6, both Z_{sem} and Z_{shape} contribute to improved segmentation performance, but using Z_{sem} alone yields better results than using Z_{shape} alone.

| Method | Backbone | Setting | PASCAL-5 ⁱ | | | | | | COCO-20 ⁱ | | | | | |
|-------------|----------|-----------|-----------------------|----------------|----------------|----------------|-------------|-------------|----------------------|-----------------|-----------------|-----------------|-------------|-------------|
| | | | 5 ⁰ | 5 ¹ | 5 ² | 5 ³ | mIoU | FBIoU | 20 ⁰ | 20 ¹ | 20 ² | 20 ³ | mIoU | FBIoU |
| FWB [38] | ResNet | 1-shot | 51.3 | 64.5 | 56.7 | 52.2 | 56.2 | — | 17.0 | 18.0 | 21.0 | 28.9 | 21.2 | — |
| DAN [52] | ResNet | 1-shot | 54.7 | 68.6 | 57.8 | 51.6 | 58.2 | 71.9 | — | — | — | — | 24.4 | 62.3 |
| PFENet [51] | ResNet | 1-shot | 60.5 | 69.4 | 54.4 | 55.9 | 60.1 | 72.9 | 36.8 | 41.8 | 38.7 | 36.7 | 38.5 | 63.0 |
| HSNet [37] | ResNet | 1-shot | 67.3 | 72.3 | 62.0 | 63.1 | 66.2 | 77.6 | 37.2 | 44.1 | 42.4 | 41.3 | 41.2 | 69.1 |
| SPNet [54] | ResNet | zero-shot | 23.8 | 17.0 | 14.1 | 18.3 | 18.3 | 44.3 | — | — | — | — | — | — |
| ZS3Net [4] | ResNet | zero-shot | 40.8 | 39.4 | 39.3 | 33.6 | 38.3 | 57.7 | 18.8 | 20.1 | 24.8 | 20.5 | 21.1 | 55.1 |
| LSeg [25] | ResNet | zero-shot | 52.8 | 53.8 | 44.4 | 38.5 | 47.4 | 64.1 | 22.1 | 25.1 | 24.9 | 21.6 | 23.4 | 57.9 |
| Ours | DRN | zero-shot | 57.3 | 60.3 | 58.4 | 45.9 | 55.5 | 66.4 | 34.2 | 36.5 | 34.6 | 35.6 | 35.2 | 58.4 |
| LSeg [25] | ViT-L | zero-shot | 61.3 | 63.6 | 43.1 | 41.0 | 52.3 | 67.6 | 28.1 | 27.5 | 30.0 | 23.2 | 27.2 | 59.9 |
| Ours | ViT-L | zero-shot | 62.7 | 64.3 | 60.6 | 50.2 | 59.4 | 69.0 | 33.8 | 38.1 | 34.4 | 35.0 | 35.3 | 58.2 |

Table 1. The performances of SAZS and baselines evaluated on PASCAL-5ⁱ and COCO-20ⁱ

| Model | Backbone | external dataset | target dataset | PASCAL-5 ⁱ |
|--------------|----------|------------------|------------------|-----------------------|
| LSeg | ViT-L | ✗ | ✓ (seen classes) | 52.3 |
| SPNet | ResNet | ✗ | ✓ (seen classes) | 18.3 |
| ZS3Net | ResNet | ✗ | ✓ (seen classes) | 38.3 |
| LSeg | ResNet | ✗ | ✓ (seen classes) | 47.4 |
| LSeg+ | ResNet | COCO | ✗ | 59.0 |
| OpenSeg [14] | ResNet | COCO | ✗ | 60.0 |
| Ours | DRN | COCO | ✗ | 62.7 |

Table 2. The cross dataset mIoU results of our model and previous SOTA methods on PASCAL-5ⁱ.

| Model | Fusion | $\mathcal{L}_{\text{shape}}$ | 5 ⁰ | 5 ¹ | 5 ² | 5 ³ | mIoU |
|-----------|--------|------------------------------|----------------|----------------|----------------|----------------|-------------|
| SAZS | ✓ | ✓ | 62.7 | 64.3 | 60.6 | 50.2 | 59.4 |
| SAZS | ✓ | | 63.1 | 62.4 | 59.0 | 49.2 | 58.4 |
| SAZS | | ✓ | 59.7 | 63.4 | 44.3 | 42.2 | 52.4 |
| SAZS | | | 59.2 | 61.9 | 43.8 | 41.9 | 51.7 |
| LSeg [25] | | | 61.3 | 63.6 | 43.1 | 41.0 | 52.3 |

Table 3. Ablation study on PASCAL-5ⁱ (ViT backbone)

| Model | Fusion | $\mathcal{L}_{\text{shape}}$ | 20 ⁰ | 20 ¹ | 20 ² | 20 ³ | mIoU |
|-----------|--------|------------------------------|-----------------|-----------------|-----------------|-----------------|-------------|
| SAZS | ✓ | ✓ | 33.8 | 38.1 | 34.4 | 35.0 | 35.3 |
| SAZS | ✓ | | 33.3 | 39.0 | 33.9 | 32.7 | 34.7 |
| SAZS | | ✓ | 30.0 | 30.4 | 27.5 | 28.5 | 29.1 |
| SAZS | | | 26.3 | 32.0 | 26.2 | 26.2 | 27.7 |
| LSeg [25] | | | 28.1 | 27.5 | 30.0 | 23.2 | 27.2 |

Table 4. Ablation study on COCO-20ⁱ (ViT backbone)

| Model | Fusion | $\mathcal{L}_{\text{shape}}$ | 20 ⁰ | 20 ¹ | 20 ² | 20 ³ | mIoU |
|-----------|--------|------------------------------|-----------------|-----------------|-----------------|-----------------|-------------|
| SAZS | ✓ | ✓ | 34.2 | 36.5 | 34.6 | 35.6 | 35.2 |
| SAZS | ✓ | | 33.7 | 38.2 | 33.4 | 35.5 | 35.2 |
| SAZS | | ✓ | 28.4 | 27.6 | 25.4 | 25.1 | 26.6 |
| SAZS | | | 24.2 | 28.5 | 24.4 | 23.3 | 25.1 |
| LSeg [25] | | | 22.1 | 25.1 | 24.9 | 21.6 | 23.4 |

Table 5. Ablation study on COCO-20ⁱ (DRN backbone)

While the segmentation performance obtained by combining the two is slightly higher than that achieved with Z_{sem} alone, using both requires fine-tuning the hyper-parameter λ , which can be unstable and requires additional effort.

4.6. Effects of Target Shape Compactness

In this section, we investigate the impact of shape-awareness on the performance of SAZS in zero-shot seman-

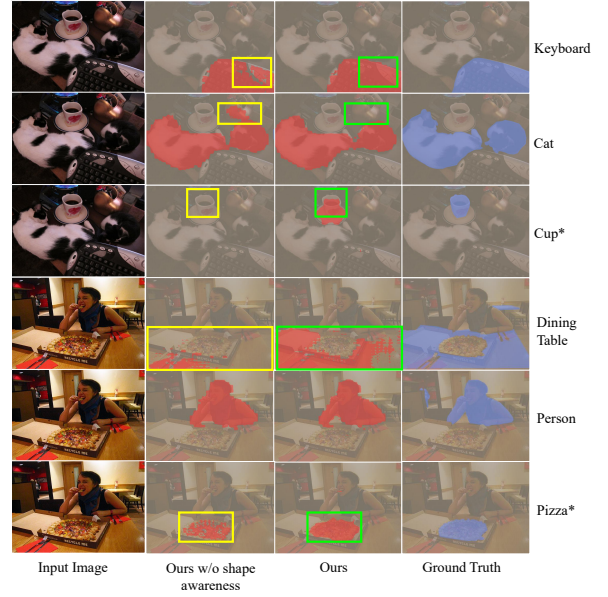


Figure 3. Qualitative results of COCO-20ⁱ. The first and last columns are the input images and the corresponding ground-truth semantic masks for different categories. The second and the third columns are the predictions by SAZS without and with shape awareness, respectively. * denotes unseen categories during training phase) and yellow boxes mark poorly segmented regions.

| Model | external dataset | Z_{shape} | Z_{sem} | PASCAL-5 ⁱ |
|-------|------------------|--------------------|------------------|-----------------------|
| SAZS | COCO | | | 58.4 |
| SAZS | COCO | ✓ | | 58.6 |
| SAZS | COCO | | ✓ | 62.7 |

Table 6. Impact of Z_{shape} and Z_{sem} Of fusion module (in the cross-dataset setting of Tabel. 2).

tic segmentation by analyzing the correlation between the mean intersection-over-union (mIoU) and the shape compactness (CO) of each category. Shape compactness, as proposed by Schick and others in 2012 [46], is a commonly used metric for measuring the similarity of superpixels to circles, which we use to characterize the shapes of objects in the input images.

For each input image in the PASCAL-5ⁱ dataset, we

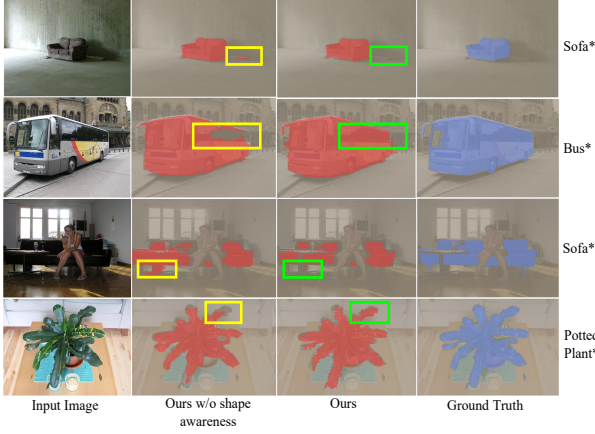


Figure 4. Qualitative comparison results of PASCAL-5ⁱ. The first and last columns are the input images and the corresponding ground-truth semantic masks for different categories. The second and third columns are the predictions by SAZS without and with shape awareness, respectively. * denotes categories (categories unseen during training phase) and yellow boxes mark poorly segmented regions.

collected the compactness (CO) metric of the ground-truth mask for the target object to describe its shape. We then calculated the variance of CO for each object category and plotted the results in Fig. 5a. The sample points in the figure represent the IoU and CO variance of each category, with the color indicating the experiment settings. This analysis aims to investigate how shape-awareness affects the SAZS’s performance on zero-shot semantic segmentation.

The results demonstrate a negative correlation between the IoU and the CO variance of a specific category (with a Pearson correlation coefficient of $r > 0.7$ and $P \leq 0.001$), and the degree of correlation is higher for SAZS than for the baselines. These findings strongly suggest that shape-awareness can improve segmentation performance when objects have more stable shapes, and that SAZS is more able to leverage shape information compared to the other baselines. The experiments were conducted on the PASCAL-5ⁱ dataset.

4.7. Effects of the Language Embedding Locality

Intuitively, distribution of language anchors in the latent feature space may largely affect vision-language alignment and thus the performance of the proposed method. Inspired by recent research [21, 26, 42], we model the distribution by the embedding locality of anchors which is defined by the mean value and standard deviation of euclidean distances in the feature space between one anchor and all other anchors.

For each category in each setting of experiments, we calculate its embedding locality and report the results collected on PASCAL-5ⁱ in Fig. 5b. The coordinates of sample points

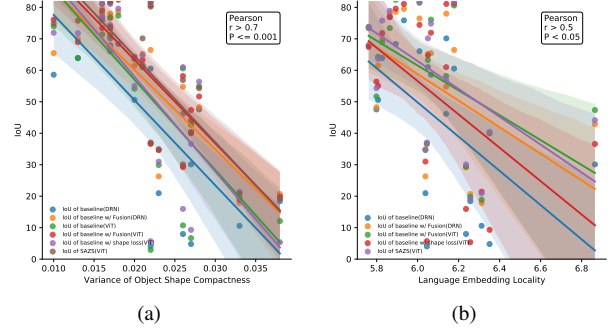


Figure 5. Correlation of CO variance (a) or mean embedding locality (b) with IoU.

represent the IoU and the embedding locality of the corresponding category while the colors of the sample points denote the experiment settings.

According to the plotted results, we observe a negative linear correlation (with Pearson correlation coefficient $r > 0.5$ and $P \leq 0.05$) between the embedding locality mean and IoU of a certain category, indicating that the closer a category is in the feature space to the others, the easier it is for the visual and text embeddings which leads to higher performances. Also, the degree of relevance of SAZS is the highest among all methods which implies that SAZS is able to better align pixel-wise visual embeddings towards the text anchors in the CLIP feature space.

5. Conclusion

In this paper, we present a novel framework for Shape-Aware Zero-Shot semantic segmentation (abbreviated as SAZS). The proposed framework leverages the rich priors contained in the feature space of a large-scale pre-trained visual-language model, while also incorporating shape-awareness through joint training on a boundary detection constraint task. This is necessary to compensate for the absence of fine-grained features in the feature space. In addition, self-supervised spectral decomposition is used to obtain feature vectors for images, which are fused with the network predictions as prior knowledge to enhance the model’s ability to perceive shapes.

Extensive experiments demonstrate the state-of-the-art performance of SAZS with significant margins over previous methods. Correlation analysis further highlights the impact of shape compactness and distribution of language anchors on the framework’s performance. Our approach effectively exploits the shape of targets and feature priors, showing the highest correlation among all compared methods and proving the novelty of the shape-aware design.

References

- [1] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. **3**
- [2] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9536–9545, 2021. **2**
- [3] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5911, 2021. **5**
- [4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. **1, 2, 3, 7**
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. **5**
- [6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. **1, 2**
- [7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. **1**
- [8] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013. **5**
- [9] Xiaoxue Chen, Tianyu Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Cerberus transformer: Joint semantic, affordance and attribute parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19649–19658, 2022. **1**
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. **3**
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. **6**
- [12] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. **6, 12**
- [13] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–510, 2021. **3**
- [14] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. **6, 7**
- [15] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020. **1, 2, 3**
- [16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. **6**
- [17] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 33:21713–21724, 2020. **3**
- [18] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. **1**
- [19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. **1**
- [20] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer. *arXiv preprint arXiv:2302.00673*, 2023. **3**
- [21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrm: modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. **8**
- [22] Naoki Kato, Toshihiko Yamasaki, and Kiyoharu Aizawa. Zero-shot semantic segmentation via variational mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. **2, 3**
- [23] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9207–9216, 2019. **2**
- [24] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007. **5**
- [25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. **1, 2, 3, 6, 7**

- [26] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 8
- [27] Pengfei Li, Beiwen Tian, Yongliang Shi, Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Toist: Task oriented instance segmentation transformer with noun-pronoun distillation. *arXiv preprint arXiv:2210.10775*, 2022. 3
- [28] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 33:10317–10327, 2020. 1, 2, 3
- [29] Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. Understanding embodied reference with touch-line transformer. *arXiv preprint arXiv:2210.05668*, 2022. 3
- [30] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 12
- [31] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30, 2017. 2
- [32] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021. 1
- [33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2
- [34] Fengmao Lv, Haiyang Liu, Yichen Wang, Jiayi Zhao, and Guowu Yang. Learning unbiased zero-shot semantic segmentation networks via transductive transfer. *IEEE Signal Processing Letters*, 27:1640–1644, 2020. 2
- [35] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 5
- [36] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3
- [37] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952, 2021. 2, 4, 7
- [38] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 622–631, 2019. 2, 7
- [39] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2693–2702, 2021. 3
- [40] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3
- [41] Trung Pham, Thanh-Toan Do, Gustavo Carneiro, Ian Reid, et al. Bayesian semantic instance segmentation in open set world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018. 1
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 8
- [43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 4, 6
- [44] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 3
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2
- [46] Alexander Schick, Mika Fischer, and Rainer Stiefelhagen. Measuring and evaluating the compactness of superpixels. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 930–934. IEEE, 2012. 7
- [47] Jianjun Shen, Siyi Lu, Ruize Qu, Hao Zhao, Yu Zhang, An Chang, Li Zhang, Wei Fu, and Zhipeng Zhang. Measuring distance from lowest boundary of rectal tumor to anal verge on ct images using pyramid attention pooling transformer. *Computers in Biology and Medicine*, 155:106675, 2023. 1
- [48] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 3
- [49] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5229–5238, 2019. 2
- [50] Beiwen Tian, Liyi Luo, Hao Zhao, and Guyue Zhou. Vibus: Data-efficient 3d scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:302–318, 2022. 1
- [51] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 7
- [52] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *European*

- Conference on Computer Vision*, pages 730–746. Springer, 2020. 2, 7
- [53] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 2
 - [54] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 2, 3, 7
 - [55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1
 - [56] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 3
 - [57] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 4, 6
 - [58] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020. 1
 - [59] Pengju Zhang, Yihong Wu, and Jiagang Zhu. Semi-global shape-aware network. *arXiv preprint arXiv:2012.09372*, 2020. 2
 - [60] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Pointly-supervised scene parsing with uncertainty mixture. *Computer Vision and Image Understanding*, 200:103040, 2020. 1
 - [61] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, 2017. 3
 - [62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1
 - [63] Yupeng Zheng, Chengliang Zhong, Pengfei Li, Huan-ang Gao, Yuhang Zheng, Bu Jin, Ling Wang, Hao Zhao, Guyue Zhou, Qichao Zhang, et al. Steps: Joint self-supervised nighttime image enhancement and depth estimation. *arXiv preprint arXiv:2302.01334*, 2023. 1
 - [64] Leisheng Zhong, Yu Zhang, Hao Zhao, An Chang, Wenhao Xiang, Shunli Zhang, and Li Zhang. Seeing through the occluders: Robust monocular 6-dof object pose tracking via model-guided video object segmentation. *IEEE Robotics and Automation Letters*, 5(4):5159–5166, 2020. 1

6. Appendix

6.1. Per-Category Evaluation

Table 8 and Table 9 demonstrate our per-category zero-shot semantic mIoU results on COCO-20ⁱ [30] and PASCAL-5ⁱ [12], respectively. The mIoU of our proposed SAZS network structure demonstrates superior performance compared to the baseline. We also observed that certain categories often appear as small regions, such as ties, or have complicated internal structures, such as people. For these categories, textual feature guidance alone cannot provide sufficient information for semantic parsing, and the baseline without shape-awareness cannot effectively segment objects under self-supervision. However, when using a SAZS model, the mIoUs of these categories better align with the shapes of the objects than the baseline, which confirms that shape awareness indeed improves zero-shot learning.

6.2. Speed and Complexity

We conducted experiments to analyze the per-episode inference time and floating point operations per second (FLOPs) in order to demonstrate the complexity of our proposed approach. The results are summarized in Table 7 for the COCO-20ⁱ dataset. Compared to the baseline model without the fusion module, SAZS had slower inference time, but significantly better performance. Although losses, including L_{shape} , in our model did not add any time cost during inference, there is still potential for optimization in terms of inference speed and model complexity, which is exactly the direction for our future research.

6.3. More Qualitative Results

In this section, we provide additional qualitative results of our model with a ViT-L backbone on PASCAL-5ⁱ and COCO-20ⁱ datasets to demonstrate the model’s ability to perform semantic segmentation on previously unseen categories. Fig. 7 showcases the results on PASCAL-5ⁱ, where all categories are unseen in their respective fold. The images presented in the figure vary in their content and complexity, and we display different visualizations of SAZS to demonstrate its versatility.

The results presented in Fig. 7 demonstrate the efficacy of SAZS in distinguishing the target semantic objects, such as bicycle, dining table, and TV monitor, from distractors like person, dog, and keyboard. Furthermore, in Fig. 7, SAZS accurately segments multiple instances of the target object, as is the case with the train, potted plant, and TV monitor.

Overall, these results demonstrate the robustness of our model in semantically segmenting novel categories with high precision and accuracy, even in complex scenes. In this section, we present the visualization of COCO-20ⁱ in Fig. 8, which includes both seen and unseen categories. We se-

| Model | Backbone | mIoU | time(s) | FLOPS(G) |
|------------|----------|------|---------|----------|
| w/o fusion | DRN | 26.6 | 177.43 | 275.76 |
| w/o fusion | ViT-L | 29.1 | 196.95 | 345.99 |
| SAZS | DRN | 35.2 | 230.54 | 275.76 |
| SAZS | ViT-L | 35.3 | 222.52 | 345.99 |

Table 7. More quantitative results on COCO-20ⁱ.

lected 20 scene and attribute labels with different semantics and multiple objects to demonstrate the versatility of SAZS. Despite the presence of noise and complexity in the scenes, SAZS accurately recognizes novel categories that are small and intricate, as illustrated by the examples of broccoli, pottedplant, and skis in Fig. 8.

In particular, in the second image of lines 2 and 3 of Figure 1, where multiple species appear in the scene with complex shapes, SAZS performs sharp object edge segmentation to accurately distinguish broccoli, carrots, and hot dogs.

Given the diversity of the presented scenes, we believe that SAZS is precise enough to be applied to various scenarios, including open scenario understanding and intelligent service robots.

6.4. More Scatter Analysis

Fig. 6 presents additional scatterplots and corresponding Pearson analysis results for the pascal dataset. The sample points in Fig. 6 represent the IoU and CO variance of each model, and they demonstrate a negative correlation. The results indicate that our approaches, particularly those that incorporate shape-awareness, can increase the correlation between per-category IoU results and CO. For example, in the third column of Fig. 6, the Pearson correlation coefficient r of SAZS is 0.13 higher than that of the baseline.

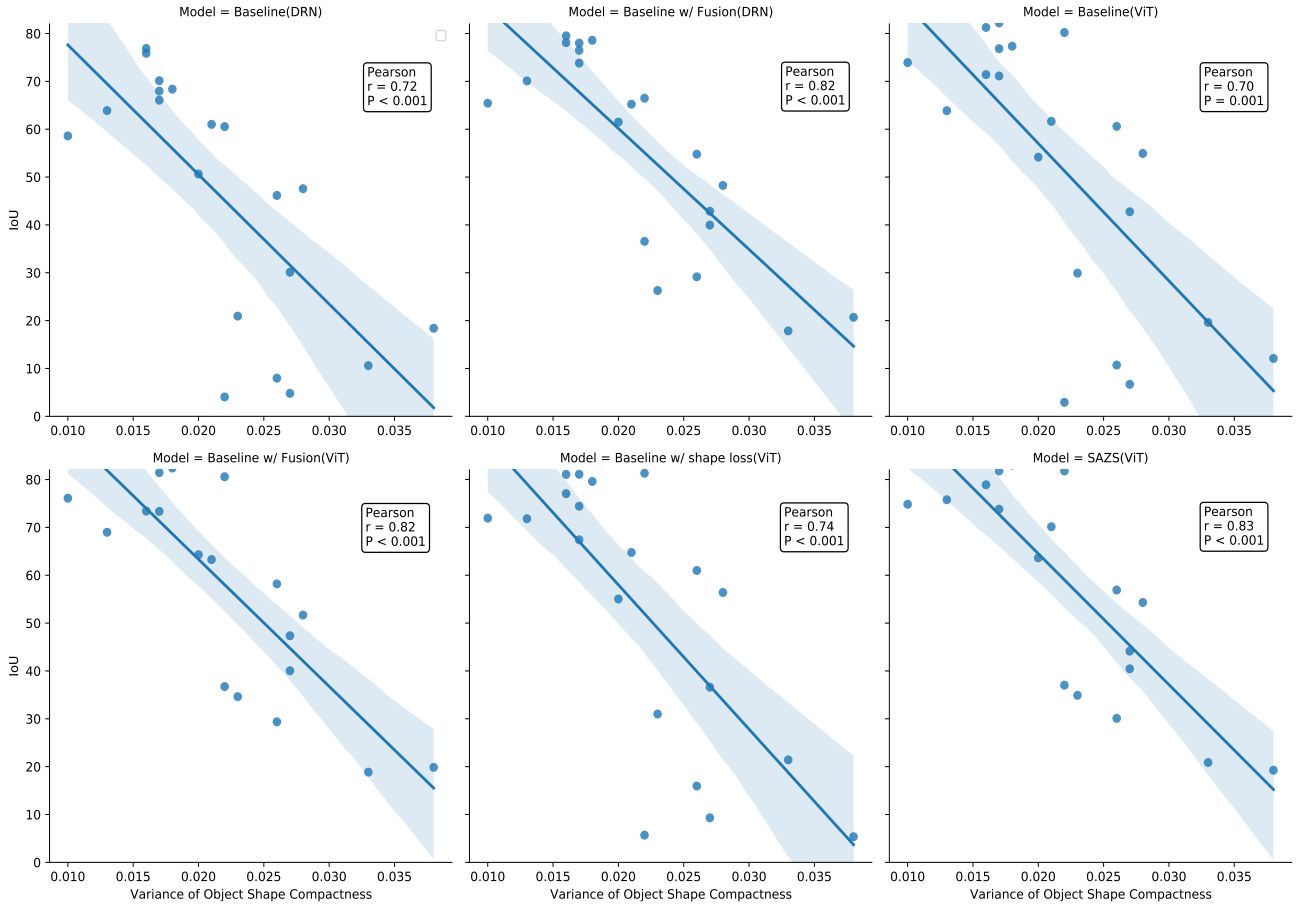
Table 8. Per-category zero-shot semantic segmentation results on COCO-20ⁱ.

| Method | Backbone | person | Bicycle | Car | motorbike | aeroplane | Bus | train | truck | boat | trafficlight | firehydrant | stopsign | parkingmeter | bench | bird | cat | dog | horse | sheep | cow | mIoU |
|-------------|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|--------------|-------------|-------------|---------------|-------------|-------------|--------------|-------------|-------------|
| Baseline | DRN | 35.7 | 55.5 | 38.2 | 43.6 | 69.2 | 69.9 | 15.1 | 27.2 | 20.2 | 24.0 | 12.2 | 5.9 | 57.2 | 66.5 | 11.9 | 43.3 | 12.3 | 19.7 | 25.3 | 20.7 | 35.2 |
| Ours | DRN | 36.0 | 61.5 | 38.5 | 55.7 | 66.7 | 72.2 | 17.9 | 29.4 | 16.7 | 14.4 | 12.2 | 5.9 | 53.8 | 65.8 | 11.5 | 46.2 | 13.5 | 15.5 | 27.6 | 22.5 | 35.2 |
| Baseline | ViT-L | 35.7 | 55.1 | 32.1 | 47.2 | 75.6 | 83.5 | 16.2 | 20.3 | 16.1 | 12.4 | 12.2 | 5.9 | 60.1 | 72.2 | 12.0 | 36.3 | 11.0 | 15.8 | 25.2 | 20.7 | 34.7 |
| Ours | ViT-L | 35.7 | 56.5 | 33.4 | 48.2 | 74.7 | 83.2 | 16.2 | 25.0 | 17.6 | 13.1 | 12.1 | 7.3 | 56.4 | 71.9 | 12.3 | 35.3 | 13.8 | 17.6 | 25.3 | 21.1 | 35.3 |
| Method | Backbone | elephant | bear | zebra | giraffe | backpack | umbrella | handbag | tie | suitcase | frisbee | skis | snowboard | sportsball | kite | baseballbat | baseballglove | skateboard | surfboard | tennisracket | bottle | |
| Baseline | DRN | 25.6 | 67.0 | 23.9 | 16.5 | 64.6 | 74.5 | 35.5 | 27.5 | 55.3 | 49.7 | 10.3 | 21.1 | 41.0 | 78.8 | 28.7 | 33.5 | 18.8 | 18.2 | 12.1 | 60.5 | |
| Ours | DRN | 24.3 | 63.8 | 23.2 | 16.5 | 57.0 | 74.5 | 35.8 | 38.2 | 53.3 | 40.6 | 10.9 | 21.1 | 38.5 | 63.9 | 20.9 | 30.3 | 20.3 | 18.2 | 15.9 | 62.2 | |
| Baseline | ViT-L | 15.9 | 61.3 | 18.8 | 16.9 | 60.0 | 79.0 | 35.5 | 52.1 | 51.9 | 47.0 | 10.4 | 21.1 | 43.7 | 85.6 | 20.1 | 33.9 | 24.7 | 18.9 | 13.0 | 69.9 | |
| Ours | ViT-L | 16.2 | 57.1 | 19.3 | 16.5 | 61.0 | 78.7 | 35.5 | 53.7 | 52.1 | 43.6 | 10.3 | 21.1 | 40.8 | 78.5 | 19.8 | 32.8 | 21.4 | 18.9 | 15.0 | 69.2 | |
| Method | Backbone | wineglass | cup | fork | knife | spoon | bowl | banana | apple | sandwich | orange | broccoli | carrot | hotdog | pizza | donut | cake | chair | sofa | pottedplant | bed | |
| Baseline | DRN | 16.8 | 56.2 | 60.3 | 47.3 | 72.5 | 73.9 | 4.7 | 9.8 | 9.6 | 18.8 | 3.7 | 46.7 | 38.1 | 62.2 | 10.2 | 17.2 | 28.1 | 13.0 | 42.0 | 36.7 | |
| Ours | DRN | 15.0 | 57.0 | 49.6 | 44.8 | 73.5 | 77.1 | 5.0 | 8.6 | 9.7 | 23.4 | 3.9 | 46.4 | 37.4 | 67.7 | 10.5 | 17.2 | 37.0 | 12.0 | 49.1 | 47.0 | |
| Baseline | ViT-L | 15.4 | 62.5 | 52.3 | 38.8 | 78.8 | 79.5 | 4.8 | 8.3 | 9.8 | 16.1 | 3.7 | 42.9 | 37.4 | 60.0 | 10.5 | 25.5 | 39.6 | 11.6 | 39.1 | 41.5 | |
| Ours | ViT-L | 14.5 | 58.6 | 58.9 | 39.0 | 79.3 | 80.2 | 4.8 | 10.2 | 9.7 | 16.0 | 4.1 | 44.6 | 37.4 | 60.6 | 10.5 | 18.1 | 36.7 | 11.4 | 46.7 | 47.3 | |
| Method | Backbone | diningtable | toilet | tvmonitor | laptop | mouse | remote | keyboard | cellphone | microwave | oven | toaster | sink | refrigerator | book | clock | vase | scissors | teddybear | hairdrier | toothbrush | |
| Baseline | DRN | 39.6 | 55.9 | 44.6 | 74.3 | 77.7 | 70.6 | 14.0 | 32.5 | 13.1 | 12.0 | 8.1 | 25.9 | 24.0 | 34.3 | 43.7 | 25.7 | 37.3 | 11.4 | 30.9 | 33.5 | |
| Ours | DRN | 54.7 | 44.9 | 47.6 | 60.5 | 70.5 | 64.0 | 16.7 | 34.5 | 13.8 | 11.2 | 7.2 | 26.2 | 24.8 | 42.6 | 43.6 | 26.0 | 47.6 | 15.5 | 32.6 | 28.4 | |
| Baseline | ViT-L | 39.2 | 32.6 | 45.7 | 70.4 | 79.5 | 67.0 | 14.0 | 20.6 | 13.4 | 10.7 | 6.4 | 26.1 | 24.1 | 37.7 | 41.7 | 25.7 | 32.6 | 11.9 | 27.4 | 26.8 | |
| Ours | ViT-L | 43.9 | 38.2 | 53.1 | 77.0 | 78.9 | 71.7 | 14.0 | 16.4 | 13.1 | 10.8 | 6.2 | 26.9 | 27.9 | 40.8 | 41.8 | 26.8 | 41.9 | 12.1 | 29.8 | 28.1 | |

Table 9. Per-category zero-shot semantic segmentation results on PASCAL-5ⁱ.

| Method | Backbone | Aeroplane | Bicycle | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow | mIoU | FBIoU |
|-------------|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Baseline | DRN | 58.6 | 20.9 | 68.4 | 50.6 | 46.2 | 76.9 | 47.6 | 70.1 | 10.6 | 75.9 | 45.5 | 61.7 |
| Ours | DRN | 65.4 | 26.3 | 78.6 | 61.5 | 54.8 | 78.1 | 48.3 | 78.0 | 17.9 | 79.5 | 55.5 | 66.4 |
| Baseline | ViT/L | 76.1 | 34.6 | 82.4 | 64.3 | 58.2 | 73.4 | 51.7 | 84.7 | 18.9 | 83.1 | 58.4 | 68.3 |
| Ours | ViT/L | 74.8 | 34.9 | 83.0 | 63.6 | 56.9 | 78.9 | 54.3 | 84.0 | 20.9 | 83.2 | 59.4 | 69.0 |

| Method | Backbone | Diningtable | Dog | Horse | Motorbike | Person | Pottedplant | Sheep | Sofa | Train | Tvmonitor | | |
|-------------|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--|--|
| Baseline | DRN | 4.8 | 66.1 | 68.0 | 61.0 | 4.1 | 18.4 | 60.5 | 30.1 | 63.9 | 8.0 | | |
| Ours | DRN | 40.0 | 76.5 | 73.8 | 65.2 | 36.6 | 20.7 | 66.5 | 42.9 | 70.1 | 29.2 | | |
| Baseline | ViT/L | 40.0 | 81.5 | 73.4 | 63.3 | 36.7 | 19.9 | 80.6 | 47.4 | 69.0 | 29.4 | | |
| Ours | ViT/L | 40.5 | 81.8 | 73.8 | 70.1 | 37.0 | 19.3 | 81.8 | 44.1 | 75.8 | 30.1 | | |

Figure 6. More scatterplots on PASCAL-5ⁱ.





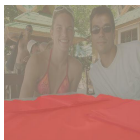


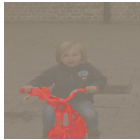


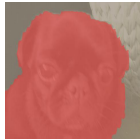
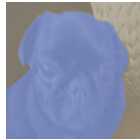
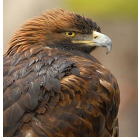
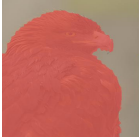
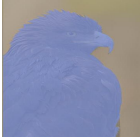
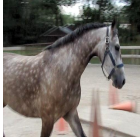
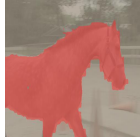





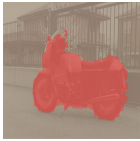
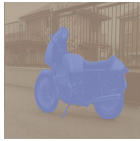



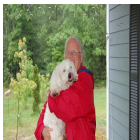

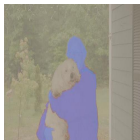



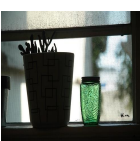
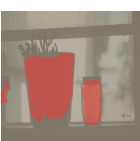
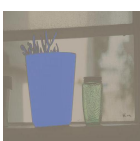

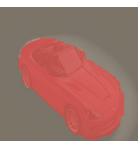

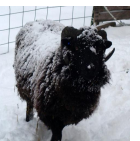

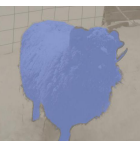
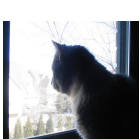
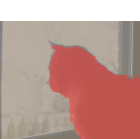




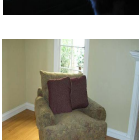
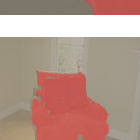
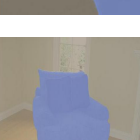

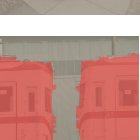


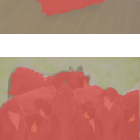
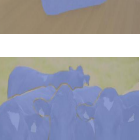

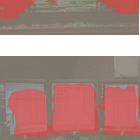

| | Image | Ours | GT | Image | Ours | GT | |
|-----------|---|---|---|---|--|---|-------------|
| Aeroplane |  |  |  |  |  |  | Diningtable |
| Bicycle |  |  |  |  |  |  | Dog |
| Bird |  |  |  |  |  |  | Horse |
| Boat |  |  |  |  |  |  | Motorbike |
| Bottle |  |  |  |  |  |  | Person |
| Bus |  |  |  |  |  |  | Pottedplant |
| Car |  |  |  |  |  |  | Sheep |
| Cat |  |  |  |  |  |  | sofa |
| Chair |  |  |  |  |  |  | Train |
| Cow |  |  |  |  |  |  | Tvmonitor |

Figure 7. More qualitative results on PASCAL-5ⁱ.

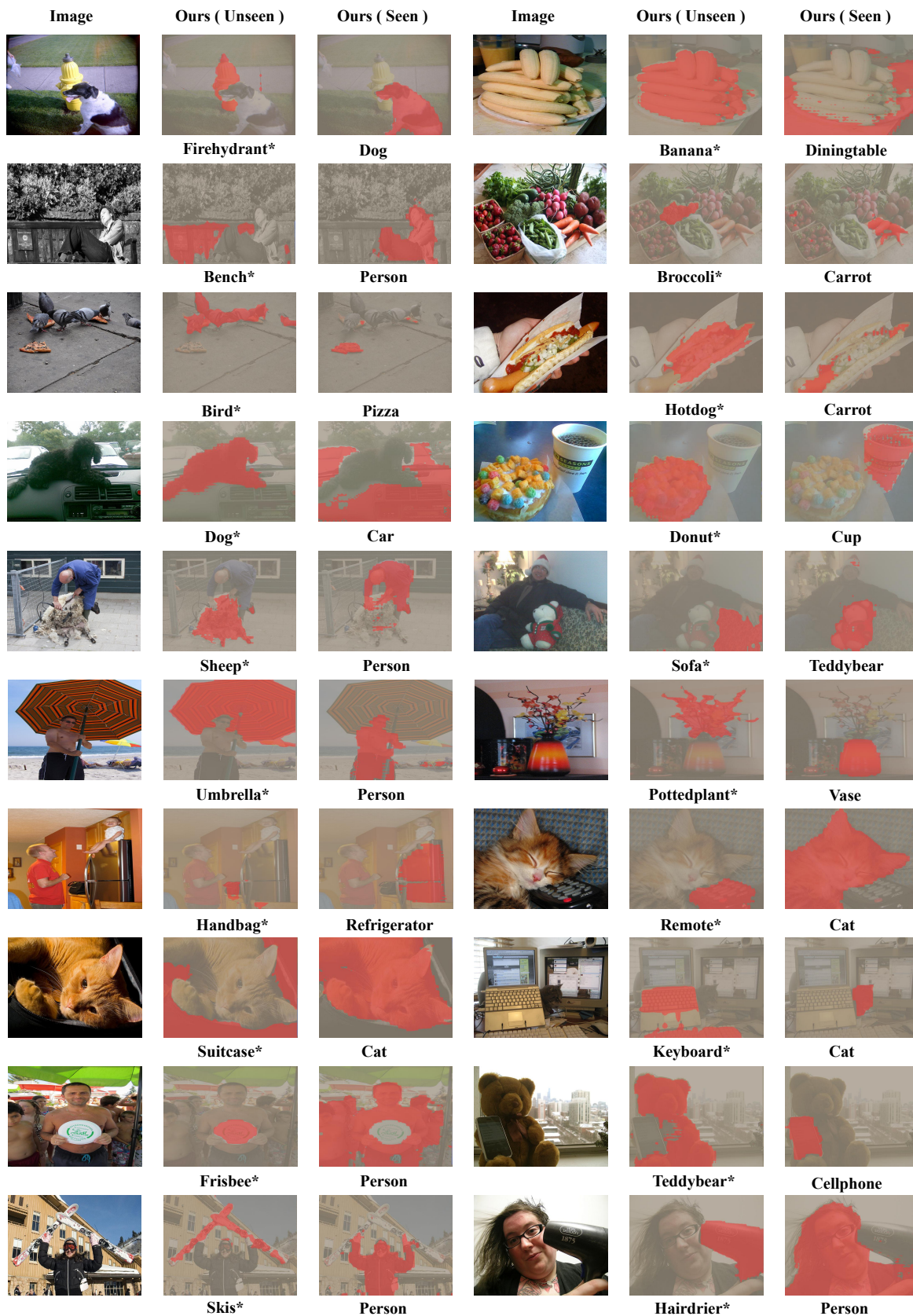


Figure 8. More qualitative results on COCO-20ⁱ.