# Zero-Shot Semantic Segmentation via Variational Mapping

Naoki Kato, Toshihiko Yamasaki, Kiyoharu Aizawa
The University of Tokyo
{kato, yamasaki, aizawa}@hal.t.u-tokyo.ac.jp

## Abstract

*We have witnessed the explosive success of deep neural networks (DNNs). However, DNNs typically assume a large amount of training data, and this is not always available in practical scenarios. In this paper, we present zero-shot semantic segmentation, where a model that has never seen the target class during training. For this purpose, we propose variational mapping, which facilitates effective learning by mapping the class label embedding vectors from the semantic space to the visual space. Experimental results using Pascal VOC 2012 show that our proposed method can achieve a mean intersection over union (mIoU) of 42.2, and we believe that this can serve as a baseline for similar research in the future.*

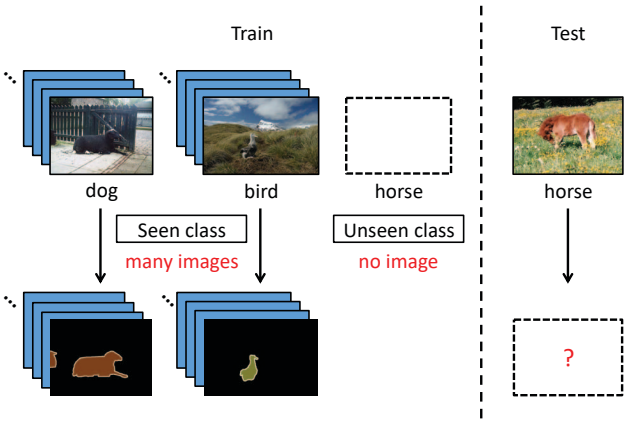


Figure 1: Overview of zero-shot semantic segmentation. There is no overlap between the seen classes in the training set and the unseen classes in the test set.

## 1. Introduction

Deep neural networks (DNNs) have demonstrated considerable success in the field of image recognition [15, 11, 28], and since then they have been used for various tasks. Although most tasks require a large amount of data, there are some cases where typical DNNs cannot be used owing to a lack of data in the real world. Few-shot learning and zero-shot learning are useful for tasks that do not have enough data. The goal of few-shot learning is to solve a task with few samples of the target class. The major approach to few-shot learning is an efficient method called meta-learning, which avoids overfitting. The goal of zero-shot learning is to solve a task with no samples of the target (unseen) class, but with semantic information regarding the seen and unseen classes. The type of semantic information depends on the dataset, but in most cases we use class label embeddings—for example, word embeddings of the classes, or attributes such as color and shape. For zero-shot learning, many researchers focus on how to use class label embeddings in the semantic space effectively.

Several studies have considered few- and zero-shot learning for solving image classification tasks. Recently, however, focus has shifted to more difficult tasks, such as zero-shot detection [5, 3] and few-shot semantic segmentation [26, 23, 7, 13]. In this paper, we undertake zero-shot semantic segmentation. This task involves performing semantic segmentation for unseen classes that do not exist at the time of training. Instead of image-mask pairs, we can use class label embeddings of the seen and unseen classes. An overview of the task is shown in Fig. 1. In this paper, we propose a method for solving this task, and we compare our method with some baseline methods. Taking into consideration previous research on zero-shot classification and few-shot semantic segmentation, our proposed method contains a two-branched approach: a segmentation branch, and a conditioning branch. We also compare the results of the experiments to few-shot semantic segmentation using the existing method.

The main contributions of this work are threefold.

1) We set a new benchmark for zero-shot semantic segmentation, inspired by previous work on few-shot semantic segmentation.

2) We show that a meta-learning approach using a two-branched architecture can be effectively applied to zero-shot semantic segmentation.

3) We propose a novel method that uses variational mapping and compare it to baseline methods.

The rest of the paper is organized as follows. In Section 2, we discuss related work on zero-shot learning and image segmentation. In Section 3, we define the problem settings and notations. In Section 4, we describe our proposed method in detail. In Section 5, we present the datasets, evaluation metric, baseline methods, and experimental results. In Section 6, we present our conclusions.

## 2. Related Work

### 2.1. Zero-Shot Learning

Several methods of zero-shot learning have been proposed in computer vision. According to [33], they are classified into 4 types: learning linear compatibility, learning nonlinear compatibility, learning intermediate attribute classifiers, and hybrid models. Models for learning linear compatibility [9, 1, 2, 24] use distance learning between visual features extracted from images by the encoders and class label embeddings for each class. Learning nonlinear compatibility includes nonlinearity in distance learning—for example, a method using the *max* function [31], and a method using the *tanh* function [29]. [16] proposes training intermediate attribute classifiers for each attribute such as a color or shape, and performing predictions with the combination of the classifiers. Hybrid model methods normally train the model using the seen classes, and then obtain prediction results for the unseen classes. This prediction result presents the ratio of what class of the seen classes is close to the unseen class, and regards it as class embedding for the query image of the unseen class. There are three patterns of feature alignment: in the latent space [35], in the semantic space [21], and in the visual space [4]. These methods achieve good results with zero-shot classification.

According to our review, there are two tasks similar to zero-shot semantic segmentation. Naha et al. [20] proposed zero-shot figure-ground segmentation, which is pixel-wise binary classification of either the foreground or background. They propose a three-step method: transferring from the seen class to the unseen class using the approach of a hybrid model method, as described above; refinement with logistic regression for each pixel, as with self-supervised learning; and a graph cut. However, it is necessary to perform logistic regression at the time of testing, and this is very time-consuming. The method is also limited to binary pixel-wise classification; multi-way segmentation is not possible with this method. Zhao et al. [36] propose open vocabulary scene parsing. They use the graph structure of WordNet [19] for jointly embedding vocabulary concepts and image pixel features, and perform multi-label semantic segmentation in consideration of the relation between hypernyms and hyponyms. For example, they can predict "furniture" from the unseen classes, using "table" or "sofa" from the seen classes. The main purpose of this method is not zero-shot learning, however. Rather, it is designed to extend the label to the hypernym or hyponym of the concept for the label, and it is premised on creating a word concept hierarchy from WordNet, etc. Therefore, the graph creation cost is considerable, and the created graph is sparse when there are few seen classes.

### 2.2. Image Segmentation

Semantic segmentation is a task of pixel-level object categorization of an input image. Many researchers use a convolutional neural network (CNN)-based method for this task, and our architecture is based on classification with a fully convolutional network (FCN) [17]. All classes of a so-called query image at testing time are the same as at training time. In addition, there are some studies on segmentation referring expressions like a keyword or sentence [27, 14, 12, 18]. However, they need various expression at training time.

Few-shot semantic segmentation is a relatively new task. In this task, only a few samples (image-mask pairs) of the same class as a query image can be used at testing time. Shaban et al. [26] first researched this task with an architecture called one-shot learning for semantic segmentation (OSLSM). [26] and all of the other previous studies [23, 7, 13] use metric-based meta-learning with a two-branched approach on K-shot one-way semantic segmentation. Note that one-way semantic segmentation tasks are binary classifications (Class 1 and the "background"), and they are so-called because the "background" label is not counted for K-way. In addition, [7] proposed two-way (Class 1, Class 2, and the "background") semantic segmentation.

### 2.3. Zero-Shot Semantic Segmentation

One work on zero-shot semantic segmentation have very recently appeared [32]. They propose a model which consists of visual-semantic embedding module that encodes images in the word embedding space and a semantic projection layer that produces class probabilities. We plan to investigate similarities and differences between our method and [32] in future work.

## 3. Problem Formulation

In this study, we tackle zero-shot one-way semantic segmentation. As such, we handle binary semantic masks. At the time of training, we use the training set $S = \{x_i, y_i(l_i), w(l_i)\}_{i=1}^{N_S}$, where for $N_S$ samples $x_i$ is the input image, $y_i(l_i)$ is the binary semantic segmentation mask, and $w_i(l_i)$ is the class label embedding, which represents the word vector of the seen class label $l_i \in L_s$ by GloVe [22]. We train the model to make the prediction mask $\hat{y_i}(l_i)$ closer to the ground-truth mask $Y_i(l_i)$ given the image $x_i$ and the word vector $w_i$. At the time of testing, we use the test set

$U = \{x_j, y_j(l_j), w(l_j)\}_{j=1}^{N_U}$, which contains $N_U$ samples related to the unseen class label $l_j \in L_u$. The goal of this task is to obtain a model that accurately predicts the semantic mask $\hat{y_j}(l_j)$ when given the image $x_j$ and the word vector $w_j$.

Note that there is no overlap between seen classes and unseen classes, i.e., $L_s \cap L_u = \emptyset$. This is the most important difference from typical semantic image segmentation, where the classes in the training set and the classes in the test set are identical. Therefore, existing segmentation methods for typical semantic segmentation, such as an FCN [17], are not effective for zero-shot semantic segmentation.

# 4. Proposed Method

## 4.1. Variational Mapping

With few-shot semantic segmentation [26, 23, 7, 13], an input image is called a query image, and a few images with annotations of the same class as the query image are called support images. The goal of this task is to predict the segmentation mask of the query image using the feature vector extracted from the support images. When training and testing, both the query images and support images are randomly sampled from the training set or the test set, respectively. Therefore, the model can learn the diverse conditions based on the support images, which are sampled many times during training.

With zero-shot semantic segmentation, by contrast, there are only class label embeddings, rather than a few images with annotations. Nevertheless, it is important to extract semantic feature maps from them, even though the extracted feature maps are not diverse. This is because the number of conditions is equal to the number of the types of these embeddings, which is also the same as the number of classes in the dataset. Therefore, the model cannot learn the diverse conditions as well as with few-shot semantic segmentation. To train the model with the diverse conditions, we propose variational mapping from the semantic space to the visual space. Variational mapping is composed of two parallel fully connected layers. When inputting a class label embedding, one of these layers is used to obtain a mean vector $\mu$, which is the same as typical mapping from the semantic space to the visual space. The other layer is used to calculate the variance vector $\sigma$, which is initialized with zero values. Then, sampling from the normal distribution $\mathcal{N}(\mu, \sigma)$, we obtain the semantic feature map $z$ in the visual space. This feature is exactly the condition with diversity.

## 4.2. Distance Metric

Some previous studies on few/zero-shot learning focus on calculating the distance between an image feature map and a semantic feature map. According to Sung et al. [30], fixed pre-specified distance metrics, such as the Euclidean or cosine distance, assume that features are solely compared element-wise. Hence, they are limited by the extent to which the feature embedding networks generate inadequately discriminate representations. Rather than fixed metrics, they propose a network with deep learning for a non-linear similarity metric jointly with embedding. In reference to their network, we adopt feature concatenation with the following embedding module for the distance metric. Specifically, a semantic feature map from variational mapping is spatially tiled and concatenated in depth with a deep image feature map extracted from the encoder. Then, we apply two $1 \times 1$ convolutions with a rectified linear unit (ReLU) function between the two convolutions to the concatenated feature in the decoder. Tiling a semantic feature map and $1 \times 1$ convolutions differs from the classification network by [30], because the output of the encoder is not a vector (tensor of rank 1), but rather a feature map (tensor of rank 3) in semantic segmentation.

## 4.3. Overall Architecture

We propose a meta-learning approach, where inputs are an image and a word vector (class label embedding), and the output is a predicted segmentation mask. An overview is shown in Fig. 2. The architecture has two branches: a conditioning branch, and a segmentation branch. In the conditioning branch, a 300-dimensional word embedding $w(l)$ of the class label $l$ is input to two parallel fully connected layers (denoted $\phi$ and $\psi$). A 128-dimensional mean vector $\mu$ and a 128-dimensional variance vector $\sigma$ are obtained. Then, a 128-dimensional semantic feature map $z$ is sampled from the normal distribution as Eq. (1). Finally, variational mapping is performed from the semantic space to the visual space.

$$z \sim \mathcal{N}(\mu, \sigma) = \mathcal{N}(\phi(w(l)), \psi(w(l))) \tag{1}$$

In the segmentation branch, by contrast, the encoder $E$ is based on the VGG-16 [28] architecture, and we remove all three fully connected layers from the original one. Following [26, 23, 7, 13], the weight of the encoder is initialized and pre-trained on ImageNet [6]. The encoder $E$ outputs a deep image feature map $m$ with 256 channels. As explained above, the image feature map $m$ and the semantic feature map $z$ with spatial tiling are concatenated in the middle of the segmentation branch. The decoder $D$ is composed two $1 \times 1$ convolutions and a transposed convolution upsampled to the original size. Note that there is a ReLU function between two convolutions, and it learns the non-linear distance in a data-driven way. Finally, the decoder $D$ receives the concatenated feature map, and outputs a predicted segmentation mask $\hat{y}(l)$ of the same size as the input image $x$. The whole architecture is expressed as Eq. (2):

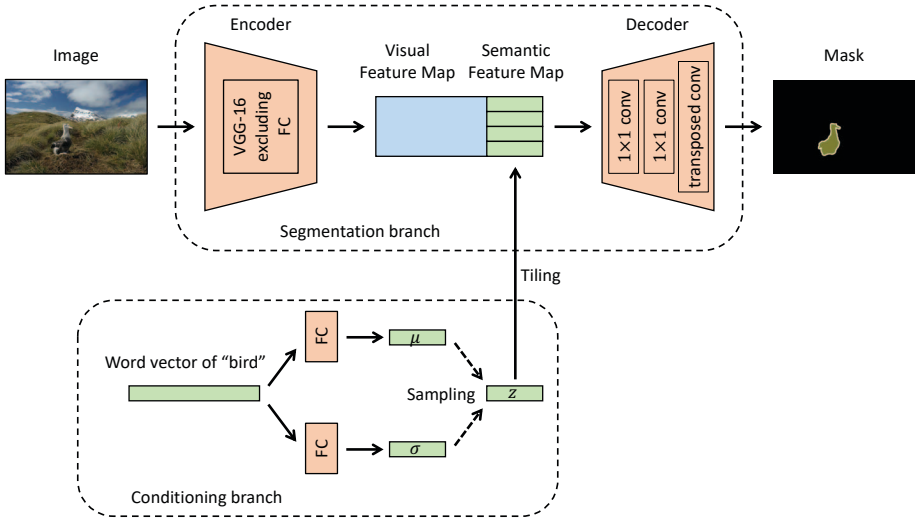


Figure 2: Overview of our two-branched architecture. The conditioning branch receives a word embedding of the class label as input, and outputs the semantic feature map. The segmentation branch receives an image as input, and outputs a predicted segmentation mask that is the same size as the input image.

$$\hat{y}(l) = D(m \oplus \mathrm{tile}(z)) = D(E(x) \oplus \mathrm{tile}(z)) \quad (2)$$

We use the pixel-wise cross entropy loss at the time of training. The entire architecture is an end-to-end trainable network for fast training.

# 5. Experiments

## 5.1. Datasets and Metric

We used PASCAL-$5^i$ (i = 0, 1, 2, 3) developed by Shaban et al. [26]. These datasets are based on the combination of images and annotations from PASCAL VOC 2012 [8][1] and its extra annotations from SDS [10][2]. The PASCAL VOC validation set was regarded as the test set, and the others as the training set. This dataset is standard for semantic segmentation. In addition, for few/zero-shot tasks, 5 classes were sampled from the 20 classes in this combined dataset and considered as the test label set $L_U = \{l_{4i+1}, ..., l_{4i+5}\}$. The remaining 15 classes were considered as the training label set $L_S$ in PASCAL-$5^i$. The names of the unseen labels (test labels) are shown in Table 1. Each PASCAL-$5^i$ contained image-annotation pairs corresponding to the split of the labels. We conducted four-fold cross-validation testing on PASCAL-$5^i$.

The class label embeddings we used were 300-dimensional word embedding vectors by GloVe [22][3] pre-trained on Common Crawl with 840B tokens. The word embeddings of the two labels whose names contain multiple words, viz., "potted plant" and "tv/monitor", are regarded as the mean vector of each word in the label name, e.g. averaging the word vector of "tv" and that of "monitor".

Table 1: Unseen classes for a four-fold cross-validation test.

| Dataset | Unseen classes |
|---|---|
| PASCAL-$5^0$ | aeroplane, bicycle, bird, boat, bottle |
| PASCAL-$5^1$ | bus, car, cat, chair, cow |
| PASCAL-$5^2$ | diningtable, dog, horse, motorbike, person |
| PASCAL-$5^3$ | potted plant, sheep, sofa, train, tv/monitor |

As a metric for the experiments, we adopted a conventional method that calculates the mean intersection over union (mIoU) of the test sets in PASCAL-$5^i$. In addition, we calculated the mean value of all the mIoU scores for each of the four tests.

## 5.2. Baselines

We compared our proposed method to the following five baseline methods in the PASCAL-$5^i$ dataset. All the baseline methods use VGG-16 [28] as backbones for their network architectures, as does our method.

**Foreground-Background (FG-BG)**
We remove the conditioning branch from our method. We do not perform feature concatenation, nor do we use the class label embeddings. This is the most naive baseline method.

[1] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/
[2] https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sds/
[3] https://nlp.stanford.edu/projects/glove/

Table 2: Results (mIoU) of zero-shot one-way semantic segmentation on PASCAL-$5^i$. Note that testing with the few-shot setting differs from the zero-shot setting.

| Methods | PASCAL-$5^0$ | PASCAL-$5^1$ | PASCAL-$5^2$ | PASCAL-$5^3$ | Mean |
|---|---|---|---|---|---|
| FG-BG | 29.6 | 38.7 | 38.1 | 32.8 | 34.8 |
| CosSim | 31.7 | 48.8 | 41.1 | 28.2 | 37.4 |
| CosSim + Mapping | 34.5 | 52.0 | **44.1** | 33.3 | 41.0 |
| Concat | 36.9 | 50.0 | 40.7 | 34.8 | 40.6 |
| Concat + Mapping | 38.7 | 50.3 | 40.0 | 31.8 | 40.2 |
| Proposed Method | **39.6** | **52.6** | 41.0 | **35.6** | **42.2** |
| FG-BG (few-shot) | 27.4 | 51.7 | 34.0 | 26.4 | 34.9 |
| OSLSM [26] (one-shot) | 33.6 | 55.3 | 40.9 | 33.5 | 40.8 |
| co-FCN [23, 34] (one-shot) | 36.7 | 50.6 | 44.9 | 32.4 | 41.1 |

**Cosine Similarity (CosSim)**
Rather than performing feature concatenation, we calculate the cosine similarity between the image feature map with 300 channels and the word embedding vector with 300 dimensions. To adjust the number of channels for the image feature map, the two $1 \times 1$ convolutions are not in the decoder, but rather in the encoder—and, in particular, not before the transposed convolution for upsampling, but before the cosine similarity operation.

**Cosine Similarity + Mapping (CosSim + Mapping)**
Instead of performing feature concatenation, we calculate the cosine similarity between the image feature map with 128 channels and the semantic feature map with 128 dimensions extracted from the word embedding vector by a fully connected layer. The $1 \times 1$ convolutions of the two layers are in the same position as the above CosSim, unlike with our proposed method.

**Concatenation (Concat)**
We do not perform any mapping of the word embedding vector from the semantic space to the visual space. We only concatenate the image feature map with 256 channels and the word embedding vector with 300 dimensions.

**Concatenation + Mapping (Concat + Mapping)**
We concatenate the image feature map with 256 channels and the semantic feature map with 128 dimensions extracted from the word embedding vector by using a fully connected layer.

### 5.3. Results

We implemented all the experiments on PASCAL-$5^i$ using PyTorch[4]. All models, including our proposed method and the baseline methods, were trained and optimized with SGD [25] with a learning rate of $10^{-5}$, momentum of 0.9, and a weight decay of 0.0005.

The experimental results are shown in Table 2. Our proposed method achieved the best mIoU scores on three of the four datasets. Moreover, from the mean of the four mIoU scores, our method overwhelmed the naive baseline (FG-BG), and it had better predictions than by not using variational mapping (Concat, Concat + Mapping). Although the combination of the cosine similarity and normal mapping also achieved a relatively good mean score, there was a difference of more than one point between CosSim + Mapping and our method. Notably, the approach using both the cosine similarity and variational mapping failed to train, and therefore the result is not included in Table 2. We believe that the reason for this is that variational mapping and fixed distance metrics such as the cosine similarity are incompatible with each other.

Some previous studies on few-shot semantic segmentation [26, 23, 7, 13] used the same backbones (VGG-16 [28]), the same datasets (PASCAL-$5^i$), and the same metric (mIoU). They sampled not only a query image but also a few annotated images at the time of testing. They used 1000 examples as samples for testing, which is not the same as with zero-shot or typical semantic segmentation. However, we found that there was little difference between the mean of the mIoU scores on the four datasets by FG-BG with zero-shot semantic segmentation compared to the few-shot setting (34.9 vs. 34.8). Hence, we assume that we can adequately compare our method with zero-shot semantic segmentation to existing methods using the few-shot setting. In one-shot semantic segmentation, the reported mean of the mIoU scores by OSLSM [26] was 40.8, and by co-FCN [23, 34], it was 41.1. Our methods achieved 42.2, and therefore zero-shot semantic segmentation can be accurately predicted to the same extent with one-shot semantic segmentation.

Fig. 3 shows some examples of the predicted qualitative results on PASCAL-$5^i$ without using the class label embed-

[4] https://pytorch.org/

Table 3: Detailed results of the pixel rate for all samples of "aeroplane" in the test set of PASCAL-$5^0$ (average IoU of "aeroplane": 58.0).

| | | Actual class | |
|---|---|---|---|
| | | background | aeroplane |
| Predicted class | background | 84.3 | 2.7 |
| | aeroplane | 3.9 | 9.1 |

Table 4: Detailed results of the pixel rate for all samples of "bottle" in the test set of PASCAL-$5^0$ (average IoU of "bottle": 33.1).

| | | Actual class | |
|---|---|---|---|
| | | background | bottle |
| Predicted class | background | 73.2 | 14.9 |
| | bottle | 3.1 | 8.9 |

dings (FG-BG), by directly using word embeddings of the class labels (Concat), and by using the semantic feature map with variational mapping (Ours). When utilizing class label embeddings, we can predict a better segmentation mask covering the object region to the edge, although in the training set there is no image of the same class as the input image.

However, there were some failure cases. We show two examples in Fig. 4. It was difficult to predict the segmentation mask of a small object in the input image, even with conventional semantic segmentation, because the texture of the object was difficult to recognize. This was especially challenging with zero-shot semantic segmentation, where there was no prior visual hint for recognizing the object. Hence, there is still room for improvement.

The tendency of the prediction differs depending on the class. We calculated the confusion matrices of the each ratio of TP, FP, FN, and TN to the sum of them for each class. The results for "aeroplane" are shown in Table 3, and those for "bottle" are shown in Table 4. Regarding the rate of FN, which refers to the number of pixels predicted as "background" that are actually in the object region, the value for "bottle" was much higher than that for "aeroplane." While both of these two classes are in the test set in PASCAL-$5^0$, it might be more difficult to predict the segmentation mask for "bottle" than for "aeroplane." This is because attributes of the object and background surrounding it in the images of "aeroplane" are similar to those of "car" and "train" in the training set, whereas there are few similar situations in the images of "bottle" to those of the other classes in the training set.

## 6. Conclusions

In this paper, we introduced a new method for zero-shot semantic segmentation, one that is more difficult than existing zero-shot classification/detection. With reference to previous research on zero-shot learning and semantic image segmentation, we proposed a novel framework to solve this new task by effectively using class label embeddings. Our two-branched architecture includes variational mapping, which helps train the model with diversity in the conditioning branch, and feature concatenation with embedding, a data-driven way to obtain a deep non-linear distance metric. We demonstrated that the proposed method outperforms all of the baseline methods.

## References


[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438, 2016.

[2] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936. IEEE Computer Society, 2015.

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, volume 11205 of *Lecture Notes in Computer Science*, pages 397–414. Springer, 2018.

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336. IEEE Computer Society, 2016.

[5] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *BMVC*, page 56. BMVA Press, 2018.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[7] Nanqing Dong and Eric Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, page 79. BMVA Press, 2018.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[9] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In Christopher J. C. Burges, Leon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *NIPS*, pages 2121–2129, 2013.

[10] Bharath Hariharan, Pablo Andres Arbelaez, Ross B. Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In David J. Fleet, Tomas Pajdla, Bernt Schiele,

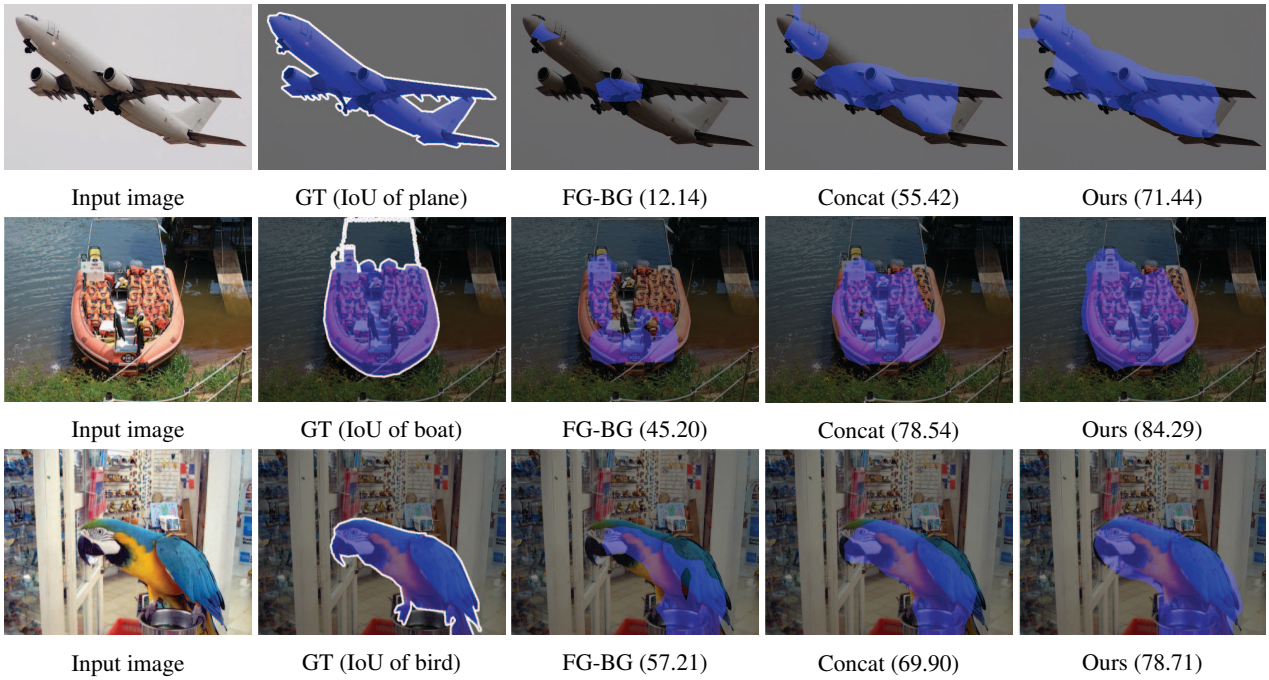


Figure 3: Qualitative results of zero-shot semantic segmentation.

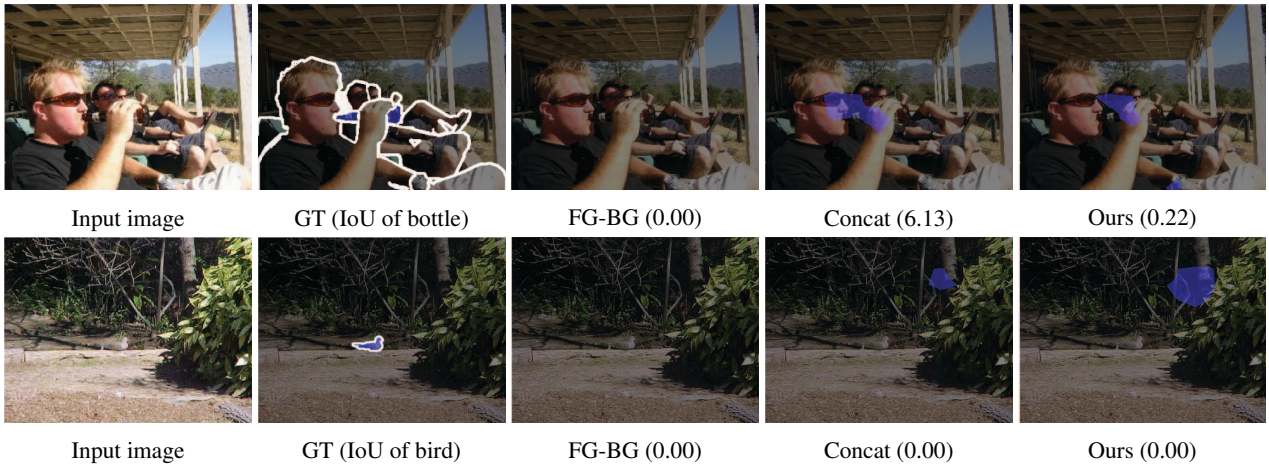


Figure 4: Failure case.

and Tinne Tuytelaars, editors, *ECCV*, volume 8695 of *Lecture Notes in Computer Science*, pages 297–312. Springer, 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[12] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, volume 9905 of *Lecture Notes in Computer Science*, pages 108–124. Springer, 2016.

[13] Tao Hu, Pengwan, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees G. M. Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI*, 2019.

[14] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *ACCV*, volume 11364 of *Lecture Notes in Computer Science*, pages 123–141. Springer, 2018.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[16] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, 2014.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, June 2015.

[18] Edgar Margffoy-Tuay, Juan C. Perez, Emilio Botero, and Pablo Arbelaez. Dynamic multimodal instance segmentation guided by natural language queries. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, volume 11215 of *Lecture Notes in Computer Science*, pages 656–672. Springer, 2018.

[19] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[20] Shujon Naha and Yang Wang. Object figure-ground segmentation using zero-shot learning. In *ICPR*, pages 2842–2847. IEEE, 2016.

[21] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014.

[22] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *EMNLP*, pages 1532–1543. ACL, 2014.

[23] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha A. Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *ICLR (Workshop)*. OpenReview.net, 2018.

[24] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2152–2161. JMLR.org, 2015.

[25] David E. Rumelhart, Geoff E. Hinton, and R. J. Wilson. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[26] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*. BMVA Press, 2017.

[27] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, volume 11210 of *Lecture Notes in Computer Science*, pages 38–54. Springer, 2018.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[29] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In Christopher J. C. Burges, Leon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *NIPS*, pages 935–943, 2013.

[30] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[31] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77. IEEE Computer Society, 2016.

[32] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, pages 8256–8265, 2019.

[33] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, pages 3077–3086. IEEE Computer Society, 2017.

[34] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *CoRR*, abs/1810.09091, 2018.

[35] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174. IEEE Computer Society, 2015.

[36] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *ICCV*, pages 2021–2029. IEEE Computer Society, 2017.