

# Semantic-Promoted Debiasing and Background Disambiguation for Zero-Shot Instance Segmentation

Shuting He<sup>1†</sup> Henghui Ding<sup>2†✉</sup> Wei Jiang<sup>1</sup>  
<sup>1</sup>Zhejiang University <sup>2</sup>Nanyang Technological University  
<https://henghuiding.github.io/D2Zero>

## Abstract

Zero-shot instance segmentation aims to detect and precisely segment objects of unseen categories without any training samples. Since the model is trained on seen categories, there is a strong bias that the model tends to classify all the objects into seen categories. Besides, there is a natural confusion between background and novel objects that have never shown up in training. These two challenges make novel objects hard to be raised in the final instance segmentation results. It is desired to rescue novel objects from background and dominated seen categories. To this end, we propose **D<sup>2</sup>Zero** with Semantic-Promoted **Deb**iasing and Background **Dis**ambiguation to enhance the performance of **Zero**-shot instance segmentation. Semantic-promoted debiasing utilizes inter-class semantic relationships to involve unseen categories in visual feature training and learns an input-conditional classifier to conduct dynamical classification based on the input image. Background disambiguation produces image-adaptive background representation to avoid mistaking novel objects for background. Extensive experiments show that we significantly outperform previous state-of-the-art methods by a large margin, e.g., **16.86%** improvement on COCO.

## 1. Introduction

Existing fully supervised instance segmentation methods [4, 23, 52] are commonly benchmarked on predefined datasets with an offline setting, where all categories are defined beforehand and learned at once, thus can neither handle novel concepts outside training datasets nor scale the model’s ability after training. Perception errors inevitably arise when applying a trained instance segmentation model to scenarios that contain novel categories. To address these challenges, zero-shot instance segmentation (ZSIS) [65] is introduced to segment instances of unseen categories with no training images but semantic information only.

<sup>†</sup>Equal contribution.

✉ Corresponding author (henghui.ding@gmail.com).

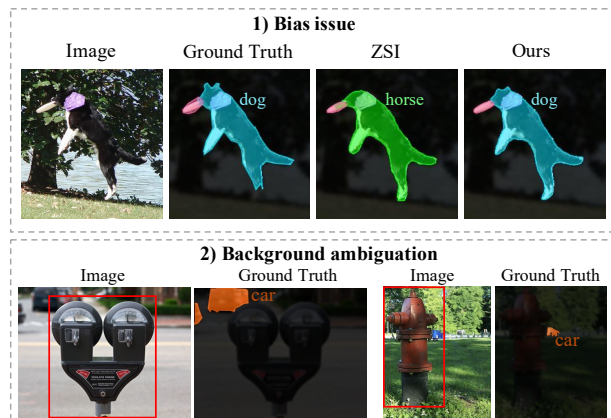


Figure 1. Two key challenges in generalized zero-shot instance segmentation. 1) Bias issue: the model tends to label novel objects with seen categories, e.g., ZSI [65] incorrectly classifies unseen class dog as training class horse. 2) Background ambiguation: objects that do not belong to any training categories are considered background, e.g., parking meter and fire hydrant.

Since scene images typically contain several objects of different categories, it is more realistic for ZSIS to segment both seen and unseen objects, which is termed Generalized ZSIS (GZSIS). In this work, we focus on two key challenges under GZSIS setting, bias issue and background ambiguation (see Figure 1), and propose **D<sup>2</sup>Zero** with semantic-promoted **Deb**iasing and background **Dis**ambiguation to enhance the performance of **Zero**-shot instance segmentation.

Bias towards seen categories imposes a significant challenge to GZSIS. Since the model is trained on data of seen categories, it tends to classify all objects into seen categories, e.g., novel object dog is labeled as seen class horse in Figure 1. Previous work ZSI [65] introduces semantic embedding to build a mapping from seen classes to unseen ones then segments novel objects by sharing instance proposals of seen group and re-labeling these proposals within unseen group. Such a “sharing” strategy brings many false positives by assigning each instance two labels. Some zero-shot semantic segmentation methods [5, 21, 35] employ a generator to synthesize fake unseen

features and fine-tune the classifier with these synthetic features. The generative way comes at the cost of forgetting some knowledge learned from seen categories and impairs the classifier’s discriminative ability of the real feature. Besides, classifier is collapsed when a new class comes in, making the generative way impractical for application. In this work, we address the bias issue from two aspects, feature extractor and classifier. Biased feature extractor mainly discriminate seen classes due to seen-only training objectives, which generalizes poorly to novel classes. We propose an unseen-constrained training objective to leverage semantic knowledge of unseen classes in visual feature learning. Specifically, we obtain semantic similarity of every seen-unseen class pair and generate a corresponding similarity-based pseudo unseen label for a seen object. Image features of seen classes are required to match the inter-class correlation with unseen classes under the supervision of pseudo unseen label, which enables the feature extractor to distinguish both seen and unseen classes.

Besides feature extractor, the bias devil also exists in the classifier. Previous zero-shot segmentation methods either use conventional fully-connected layer as classifier [5, 35] or prototypical classifier built upon semantic embeddings [61, 65]. However, these two types of classifier both have features clustered to fixed seen-class centers and do not consider the bias during inference. To address this issue, we design an input-conditional classifier based on transformer mechanism. We employ the semantic embeddings as query and visual features as key and value of transformer decoder, which bridges the semantic and visual spaces and transfers knowledge. Then the decoder outputs are employed as classifier in a prototypical way. The input-conditional classifier captures image-specific clues [40] and can better distinguish different categories of the input image. In such a way, the model learns to dynamically project semantic embeddings to input-conditional class centers, which greatly alleviates bias issue. Moreover, the input-conditional classifier establish the information interaction between visual and semantic spaces, contributing to mitigating multi-modal domain gap problem.

The background ambiguity issue is specific for zero-shot instance segmentation. In the training of instance segmentation, objects that do not belong to any training categories are considered background, *e.g.*, parking meter and hydrant in Figure 1. The model hence is likely to identify the novel objects as background, which affects the final performance a lot. To address this issue, BLC [63] and ZSI [65] propose to learn a background vector in the Region Proposal Network (RPN), which is optimized in a binary classifier of RPN. However, the binary classifier of RPN tends to overfit to seen categories and may fail to identify unseen categories [31, 59]. We experimentally find that the Transformer [51] based DETR-like model [6, 9]

can well generalize to novel categories in terms of proposal generation, thanks to its end-to-end training manner and classification-free instance proposal generation. Therefore, we collect all the foreground mask proposal produced by DETR-like model to get the global foreground mask and then apply the reverse of it on the feature map to get background prototype, which is used for background classification. Such an adaptive background prototype that updates according to input image can better capture image-specific and discriminative background visual clues, which helps to background disambiguation.

Our main contributions are summarised as follows:

- We propose an unseen constrained visual feature learning strategy to leverage semantic knowledge of unseen categories in visual feature training, which facilitates mitigating bias issue in GZSIS.
- We design an input-conditional classifier that projects semantic embedding to image-specific visual prototypes, contributing to addressing both bias issue and multi-modal domain gap issue.
- To rescue novel objects from background, we introduce an image-adaptive background representation to better capture image-specific background clues.
- We achieve new state-of-the-art performance on zero-shot instance segmentation and significantly outperform ZSI [65] by a large margin, *e.g.*, **16.86%** HM-mAP under 48/17 split on COCO.

## 2. Related Work

**Zero-Shot Image Classification** aims to classify images of unseen classes that have never shown up in training samples [18, 27, 30, 32, 33, 44]. There are two different settings: zero-shot learning (ZSL) and generalized zero-shot learning (GZSL). Under the ZSL setting [32, 44], testing images are from unseen categories only. Typical ZSL methods include classifier-based way [1, 12, 36] and instance-based way [16, 49, 58], where the former one aims to learn a visual-semantic projection to transfer knowledge and the later one aims to synthesize fake unseen samples for training. GZSL [48] aims to identify samples of both seen and unseen categories simultaneously and suffers the challenge of a strong bias towards seen categories [8]. To address the bias issue, calibration methods [7, 11, 22] and detector-based methods [3, 17, 49] are introduced. The former way aims at calibrating the classification scores of seen categories to achieve a trade-off balance between seen and unseen groups, while the detector-based way explores identifying the unseen samples as out-of-distribution and classifying these unseen samples within unseen categories.

**Zero-Shot Instance Segmentation (ZSIS)**. Fully supervised instance segmentation are extensively studied in recent years [4, 23, 52], which however are data-driven

and cannot handle unseen classes that have never shown up in training. Recently, zero-shot instance segmentation is raised by ZSI [65] to apply zero-shot learning to instance segmentation. There are two test settings: zero-shot instance segmentation (ZSIS) and generalized zero-shot instance segmentation (GZSIS), where GZSIS is more realistic since an image typically contains multiple objects of different seen/unseen categories. In this work, we mainly focus on GZSIS and address its two key challenges, bias issue, and background confusion. ZSI [65] addresses the bias issue by copying all the instances detected as seen categories and re-label these instances within unseen group, resulting in many false positives. In this work, we propose an unseen-constrained visual training strategy and input-conditional classifier to alleviate the bias issue.

**Zero-Shot Semantic Segmentation (ZSSS)** [5,54] aims to segment the image to semantic regions of seen and unseen categories, it shares some commonalities with ZSIS. Existing ZSSS methods can be divided into two ways: embedding-based methods and generative-based methods. Embedding-based methods [15,26,41,45,54,56,61] project visual and semantic features to a common space, *e.g.*, semantic, visual, or latent space, to transfer knowledge and conduct classification in this common space. Generative-based ZSSS methods [5,10,25,35] utilize a feature generator to synthesize fake features for unseen categories.

**Language-driven Segmentation** shares some similarities with ZSIS. They utilize language information to guide the segmentation, *e.g.*, referring expression segmentation [13,14,39,53,57] and open-vocabulary segmentation [19,28,34]. However, instead of following the strict zero-shot setting of excluding any unseen classes in training data, these works allow as many classes as possible to implicitly participate in model training by using image captions or referring expressions, which is however considered as information leakage in the zero-shot learning setting.

### 3. Approach

#### 3.1. Problem Formulation

In zero-shot instance segmentation, there are two non-overlapping foreground groups,  $N^s$  seen categories denoted as  $C^s$  and  $N^u$  unseen categories denoted as  $C^u$ , and a background class  $c^b$ , where  $C^s = \{c_1^s, c_2^s, \dots, c_{N^s}^s\}$  and  $C^u = \{c_1^u, c_2^u, \dots, c_{N^u}^u\}$ . Each category has a corresponding semantic embedding, denoted as  $\mathcal{A} = \{\mathbf{a}^b, \mathbf{a}_1^s, \mathbf{a}_2^s, \dots, \mathbf{a}_{N^s}^s, \mathbf{a}_1^u, \mathbf{a}_2^u, \dots, \mathbf{a}_{N^u}^u\}$ . Given an image set that contains  $N^i$  images of  $N^s$  and  $N^u$  categories, the training set  $D_{train}$  is built from training images of seen categories, *i.e.*, each training image that contains any objects of  $\{c_1^s, c_2^s, \dots, c_{N^s}^s\}$  but no object of unseen categories. According to whether considering seen classes during inference, there are two different settings, one is

ZSIS which segments objects of only the unseen categories and the other is Generalized ZSIS (GZSIS) which segments objects of both seen and unseen categories. GZSIS is more realistic since an image usually includes multiple objects and we cannot ensure there is no object of seen categories.

#### 3.2. Architecture Overview

The architecture overview of our proposed D<sup>2</sup>Zero is shown in Figure 2. We adopt ResNet-50 [24] as backbone and follow the paradigm of Mask2Former [9]. Mask2Former seamlessly converts pixel classification to mask classification with careful design of proposal embeddings  $\{\mathbf{x}_n\}_{n=1}^{N^p} \in \mathbb{R}^d$  and mask predictions  $\{\mathcal{M}_n\}_{n=1}^{N^p} \in \mathbb{R}^{H \times W}$ . Since the masks are class-agnostic, the model is endowed with the ability to generate masks for novel objects that have never shown up in training set [59]. We generate a set of prototypes as input-conditional classifier. The background prototype is generated by masked average pooling on the image feature, where the background region is decided by all the class-agnostic masks. Then we adopt seen cross-entropy loss and our proposed unseen cross-entropy loss as training objectives.

#### 3.3. Semantic-Promoted Visual Feature Debiasing

Due to the lack of unseen categories' training data, the model is trained on samples of seen categories only. As a consequence, there is a strong bias towards seen categories that the model tends to classify all the testing objects into seen categories [8]. To address the bias issue, ZSI [65] separates the classification of seen and unseen categories and labels each instance with two labels, one from seen group and the other from unseen group. This strategy, though works, sidesteps the essence of the bias problem. In this work, we explore alleviating the bias issue in zero-shot instance segmentation and attribute it to 1) biased feature extractor that focuses on producing features to discriminate seen categories and 2) biased classifier that tends to capture clues derived from training data statistics. We herein propose an unseen-constrained feature extractor and input-conditional classifier to address the biased feature extractor and biased classifier, respectively. The unseen-constrained feature extractor utilizes inter-class semantic similarities to guide the training of visual feature extractor, in which seen-unseen relationships are involved as training objective thus unseen categories can join in the visual feature training. The input-conditional classifier learns a semantic-visual alignment based on transformer and generates input-specific visual representations as classifier prototypes.

#### Unseen-Constrained Visual Feature Learning

The feature extractor trained on seen categories focuses more on features useful to discriminate seen classes, inducing a loss of information required to deal with unseen

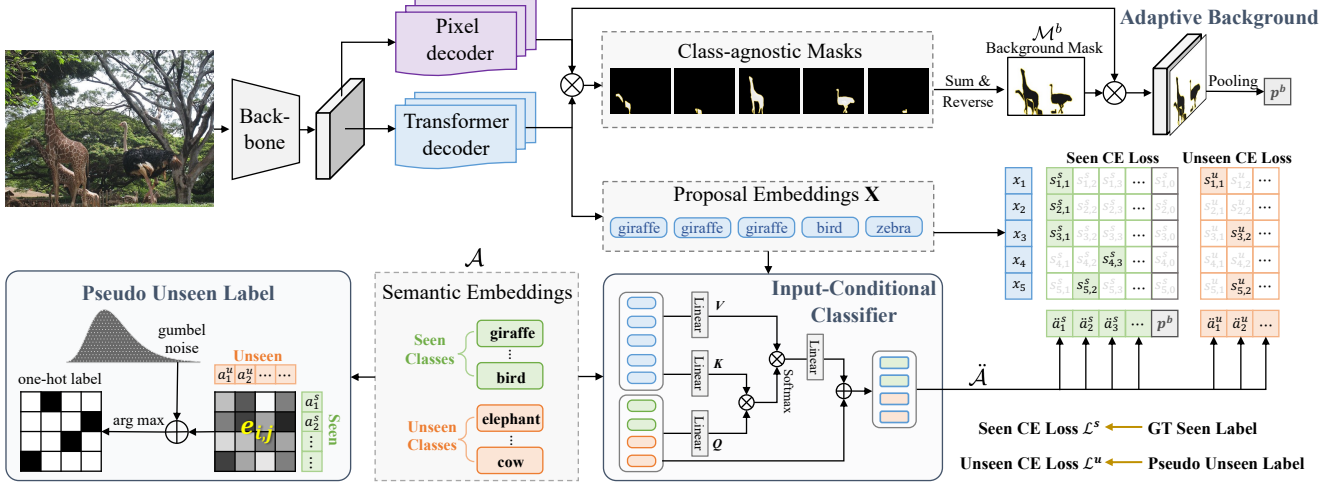


Figure 2. Framework overview of our  $D^2Zero$ . The model proposes a set of class-agnostic masks and their corresponding proposal embeddings. The proposed input-conditional classifier takes semantic embeddings and proposal embeddings as input and generates image-specific prototypes. Then we use these prototypes to classify image embeddings, under the supervision of both seen CE loss  $\mathcal{L}^s$  and unseen CE loss  $\mathcal{L}^u$ . The unseen CE loss enables unseen classes to join the training of feature extractor. We collect all the masks and produce a background mask, then apply this mask to the image feature to generate an image-adaptive background prototype for classification.

classes. To address this issue and produce features that generalize better to novel concepts, we propose to introduce semantic information of unseen categories as training guidance to constrain visual feature learning. We first generate an inter-class correlation coefficient by calculating the semantic similarity of every unseen-seen category pair,

$$e_{i,j} = \frac{\exp(\langle \mathbf{a}_i^s, \mathbf{a}_j^u \rangle / \tau)}{\sum_{k=1}^{N^u} \exp(\langle \mathbf{a}_i^s, \mathbf{a}_k^u \rangle / \tau)}, \quad (1)$$

where  $\langle, \rangle$  is cosine similarity,  $\tau$  is the temperature parameter,  $e_{i,j}$  is a soft value in the range of  $[0, 1]$  and represents correlation coefficient of the  $i$ -th seen embedding  $\mathbf{a}_i^s$  and the  $j$ -th unseen embedding  $\mathbf{a}_j^u$ , the higher  $e_{i,j}$  represents the closer relationship. For each of the  $N^s$  seen categories, there are  $N^u$  coefficients, i.e.,  $\mathbf{e}_i = \{e_{i,j}\}_{j=1}^{N^u}$ . The inter-class correlation matrix prior is then used to guide the visual feature learning. Instead of using original soft probability, we choose a pseudo unseen label for each seen object based on the coefficient  $e_{i,j}$ . Specifically, we employ Gumbel-Softmax trick [29] to form a Gumbel-Softmax distribution and transform  $\mathbf{e}_i$  to discrete variable  $\hat{\mathbf{e}}_i \in \{0, 1\}^{N^u}$

$$\hat{\mathbf{e}}_i = \text{onehot}\left(\arg \max_j [g_j + e_{i,j}]\right), \quad (2)$$

where  $g_1, \dots, g_{N^u}$  are random noise samples drawn from Gumbel  $(0, 1)$  distribution. The pseudo unseen label  $\hat{\mathbf{e}}_i$  changes with training iterations, following the rule that the larger  $e_{i,j}$  has the higher probability of  $c_j^u$  being chosen as the pseudo unseen label of  $c_i^s$ . For each proposal embedding  $\mathbf{x}_n$ , a classification score  $s_n^u$  of unseen group is obtained by

$$s_{n,j}^u = \frac{\exp(\text{MLP}(\mathbf{x}_n) \mathbf{a}_j^u / \tau)}{\sum_{k=1}^{N^u} \exp(\text{MLP}(\mathbf{x}_n) \mathbf{a}_k^u / \tau)}, \quad (3)$$

where  $s_n^u = \{s_{n,j}^u\}_{j=1}^{N^u} \in \mathbb{R}^{N^u}$ , MLP denotes Multi-layer Perceptron. The ground truth of  $s_n^u$  is  $\hat{\mathbf{e}}_{c_n}$ , where  $c_n$  denotes the ground truth seen label index of  $\mathbf{x}_n$ . The unseen cross-entropy loss  $\mathcal{L}^u$  is applied on  $s_n^u$  to have the model learn pseudo classification among unseen categories,

$$\mathcal{L}^u = -\frac{1}{N^f} \sum_{n=1}^{N^f} \sum_{j=1}^{N^u} \hat{e}_{c_n,j} \log s_{n,j}^u, \quad (4)$$

where  $N^f$  is the number of proposals with foreground labels. It's worth noting that Eq. (4) is applied on  $\mathbf{x}_n$  of foreground objects while disabled for background.

Meantime, for each proposal embedding  $\mathbf{x}_n$ , a classification score  $s_n^s$  of seen group is obtained by

$$s_{n,i}^s = \frac{\exp(\mathbf{x}_n \mathbf{a}_i^s / \tau)}{\sum_{k=0}^{N^s} \exp(\mathbf{x}_n \mathbf{a}_k^s / \tau)}, \quad (5)$$

where  $s_n^s = \{s_{n,i}^s\}_{i=0}^{N^s} \in \mathbb{R}^{(N^s+1)}$  is the classification score for the  $n$ -th proposal,  $\mathbf{a}_0^s = \mathbf{a}^b$  and  $s_{n,0}^s$  represents score of background. A cross-entropy loss is applied on  $s_n^s$  to guide the classification among seen categories,

$$\mathcal{L}^s = -\frac{1}{N^p} \sum_{n=1}^{N^p} \sum_{i=0}^{N^s} \mathbb{1}(c_n = i) \log s_{n,i}^s, \quad (6)$$

where  $\mathbb{1}(\ast)$  outputs 1 when  $\ast$  is true otherwise 0,  $c_n$  is the ground truth label of  $n$ -th proposal.  $N^p$  is the number of proposals. The overall training objective is  $\mathcal{L} = \mathcal{L}^s + \lambda \mathcal{L}^u$ .

With the unseen cross-entropy loss  $\mathcal{L}^u$ , the feature extractor is also trained under the constraints of unseen categories instead of only under the constraints of seen categories, which greatly help the feature extractor capture clues that are useful for unseen categories.

### Input-Conditional Classifier

Directly using semantic embeddings  $\mathcal{A}$  as classifier, though helps to semantically links knowledge of seen and unseen groups, makes the features clustered to fixed class centers and does not consider the bias issue in classifier. To further alleviate the bias issue in zero-shot instance segmentation, we propose an input-conditional classifier that dynamically classifies visual embeddings according to input features. As shown in Figure 2, semantic embeddings  $\mathbf{a}_i$  are employed as query  $Q$  in a transformer module, while key  $K$  and value  $V$  are concatenation of proposal embeddings, *i.e.*,  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N^p}]$ , where  $[\cdot]$  denotes concatenation operation. After transformer module, semantic-projected visual embeddings  $\tilde{\mathbf{a}}_i$  that are conditional on  $\mathbf{x}_n$  are generated. In detail, given semantic embeddings  $\mathcal{A} = [\mathbf{a}_1^s, \mathbf{a}_2^s, \dots, \mathbf{a}_{N^s}^s, \mathbf{a}_1^u, \mathbf{a}_2^u, \dots, \mathbf{a}_{N^u}^u]$ , a self-attention is first performed on  $\mathcal{A}^s$  and outputs  $\hat{\mathcal{A}}^s$ . Then cross-attention is performed as

$$Q = \mathbf{w}_Q \hat{\mathcal{A}}^s, \quad K = \mathbf{w}_K \mathbf{X}, \quad V = \mathbf{w}_V \mathbf{X},$$

$$\tilde{\mathcal{A}}^s = \text{MHA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (7)$$

where  $\mathbf{w}_Q, \mathbf{w}_K, \mathbf{w}_V$  are learnable parameters of three independent linear layers mapping inputs to the same intermediate representations of dimension  $d_k$ .  $\tilde{\mathcal{A}}^s$  is the desired image-specific semantic-visual embedding. We then update the classifier by replacing the original semantic embedding  $\mathbf{a}_j^s/\mathbf{a}_k^s$  in Eq. (5) and  $\mathbf{a}_j^u/\mathbf{a}_k^u$  in Eq. (3) to input-conditional semantic embedding  $\tilde{\mathbf{a}}_i^s/\tilde{\mathbf{a}}_k^s$  and  $\tilde{\mathbf{a}}_j^u/\tilde{\mathbf{a}}_k^u$ , respectively.

$\tilde{\mathcal{A}}$  has three main advantages over original semantic embedding  $\mathcal{A}$ . First,  $\tilde{\mathcal{A}}$  is projected from semantic space to visual space via interaction with visual proposal embedding  $\mathbf{X}$ , which helps to mitigate visual-semantic domain gap and makes the classification easier to be learned. Second,  $\tilde{\mathcal{A}}$  capture image-specific clues according to input feature and can better adaptively distinguish different categories of the input image. What’s more, the class centers by  $\tilde{\mathcal{A}}$  are input-conditional instead of fixed, thus the visual features trained with such dynamic classifier would not collapse to several fixed feature centers but tend to capture discriminative inter-class distance, which greatly helps to mitigate bias issue.

### 3.4. Image-Adaptive Background Disambiguation

There is confusion between background and unseen objects in zero-shot instance segmentation. The unseen categories do not join the training of segmentation model, which is trained to identify objects of seen categories as foreground objects and others as background, so they are easy to be mistaken for background. ZSI [65] argues that the semantic word “background” cannot represent background class and propose Background Aware RPN (BA-RPN) & Synchronized Background to use a vector learned in RPN as background representation in zero-shot

classifier. However, this learned vector is fixed after training and cannot be changed according to the input image, which limits its representation to complex backgrounds and generalization ability to novel scenarios. This background parameter is optimized in a binary classifier of RPN, which tends to overfit to seen categories and may fail to identify unseen categories [31, 59]. To address this issue, we herein propose an image-adaptive background disambiguation that adaptively generates high-quality background representation according to the input image.

Specifically, we gather all the proposed binary masks  $\{\mathcal{M}_n\}_{n=1}^{N^p}$  obtained from our model to indicate foreground region  $\mathcal{M}^f$ , *i.e.*,  $\mathcal{M}_{(x,y)}^f = \max(\mathcal{M}_{0,(x,y)}, \dots, \mathcal{M}_{N^p,(x,y)})$ , where  $(x, y)$  denotes pixel position and  $\mathcal{M}_{(x,y)}^f = 1$  represents that the pixel  $(x, y)$  belongs to foreground. It’s worth noting that we gather all the proposed masks to ensure a high recall of foreground region, which is desired to detect novel objects. The background mask  $\mathcal{M}^b$  is generated by taking the reverse of foreground mask,  $\mathcal{M}^b = 1 - \mathcal{M}^f$ . Then a Mask Average Pooling (MAP) is performed on visual feature maps to get background prototype,

$$\mathbf{p}^b = \frac{\sum_{(x,y)} \mathcal{M}_{(x,y)}^b \mathbf{F}_{(x,y)}}{\sum_{(x,y)} \mathcal{M}_{(x,y)}^b}. \quad (8)$$

We use this prototype to replace  $\mathbf{a}_0^s$  Eq. (5).  $\mathbf{p}^b$  is adaptive according to visual feature and thus can better capture image-specific and discriminative background visual clues.

**Comparison with word embedding background and learned-parameter background.** 1) Word embedding of background is either learned from large-scale text data without seeing visual data, *e.g.*, word2vec [42], or derived from text encoder trained on large-scale text-image pairs of “thing” classes, *e.g.*, CLIP [46]. Thus, the existing background word-vector cannot well represent the complex visual appearance of background. 2) ZSI [65] learns a background vector in the Region Proposal Network (RPN) and uses this vector to update the semantic embedding of background class. Such a learned vector is optimized in a binary classifier of RPN and captures some visual patterns. However, it is fixed after training and may identify novel objects as background, since the binary classifier of RPN tends to overfit to seen categories and may fail to identify unseen categories [31, 59]. 3) Our proposed image-adaptive prototype is visual feature obtained from the image background region directly and captures more useful visual clues. Compared to BA-RPN of ZSI [65] using a binary classifier, our DETR-like model can better generalize to novel categories in terms of proposing foreground instances because of its classification-free instance proposal manner. The proposed adaptive background prototype changes according to the input image and can better capture image-specific and discriminative background visual clues.

## 4. Experiments

### 4.1. Experimental Setup

**Implementation Details.** The proposed approach is implemented with the public platform Pytorch. We use ResNet-50 [24] based Mask2Former [9] to generate class-agnostic masks and corresponding proposal embeddings. All hyper-parameters are consistent with the default settings unless otherwise specified. We use CLIP [46] to extract semantic embeddings of COCO classes. Meantime, for a fair comparison with previous works, we also report our results based on word2vec [43]. Hyper-parameter  $\lambda$  and  $\tau$  are set to 0.1, 0.1, respectively. The model is optimized using Adamw with learning rate set to 0.0001, trained on 8 RTX2080Ti(12G) with batch size set to 16.

**Dataset & Training/Testing Setting.** Following ZSI [65], we use MS-COCO 2014 [38] instance segmentation dataset containing 80 classes to train and evaluate our proposed approach. Two different splits of seen and unseen categories are built to evaluate zero-shot ability. The first is 48/17 split with 48 seen categories and 17 unseen categories. The second is 65/15 split with 65 seen categories and 15 unseen categories. Training set is built from images containing seen categories only. To avoid information leakage, the images that contain any pixels of unseen categories are removed from training set, which is different from open-vocabulary setting that allows using of some unseen images [60]. In testing set, all the MS-COCO testing images that contain pixels of unseen categories are selected.

**Metrics.** Following ZSI [65], Recall@100, *i.e.*, top 100 instances, with IoU thresholds of {0.4, 0.5, 0.6} and mean Average Precision (mAP) with IoU thresholds of 0.5 are employed to report the performance. Under GZSIS setting, seen categories far outperform unseen categories and overmaster the Recall@100 and mAP. To better reveal unseen categories’ effects on overall performance, we compute the harmonic mean (HM) [55] of seen and unseen categories, where  $HM(A, B) = 2AB / (A + B)$ .

**Text Prompts.** We follow previous works [20, 46] to generate the text embeddings using prompt ensembling. For each category, we utilize multiple prompt templates and then obtain the final text embeddings via averaging.

### 4.2. Component Analysis

We conduct extensive experiments to verify the effectiveness of our proposed components in Table 1 with both 48/17 and 65/15 splits. We design our baseline by replacing the learnable classifier with text embeddings to classify image embeddings, which is similar to VILD-Text [20]. As we can see, there is a serious bias towards seen categories issue, *e.g.*, unseen mAP 7.15% is much lower than seen mAP 53.49%. In the following, we analyze our proposed component from a qualitative and quantitative perspective.

Split	$\check{A}$	$p^b$	$\mathcal{L}^u$	Seen		Unseen		HM	
				mAP	Recall	mAP	Recall	mAP	Recall
48/17	x	x	x	53.49	<b>77.52</b>	7.15	32.39	12.61	45.68
	✓	x	x	53.24	76.11	11.95	36.53	19.52	49.36
	x	✓	x	53.17	76.13	10.06	36.71	16.91	49.53
	x	x	✓	52.78	75.69	11.34	38.23	18.66	50.80
	✓	✓	✓	<b>54.42</b>	76.22	<b>15.06</b>	<b>38.38</b>	<b>23.59</b>	<b>51.06</b>
65/15	x	x	x	40.64	74.91	15.65	35.61	22.59	48.27
	✓	x	x	<b>41.26</b>	<b>75.41</b>	18.89	40.64	25.91	52.82
	x	✓	x	40.45	74.18	17.78	38.12	24.70	50.36
	x	x	✓	39.51	73.87	18.23	41.38	24.94	53.04
	✓	✓	✓	41.18	74.94	<b>20.22</b>	<b>46.01</b>	<b>27.13</b>	<b>57.01</b>

Table 1. Component Analysis of our  $D^2Zero$  under GZSIS setting.  $\check{A}$  represents input-conditional classifier.  $p^b$  is image-adaptive background.  $\mathcal{L}^u$  represents unseen CE loss.

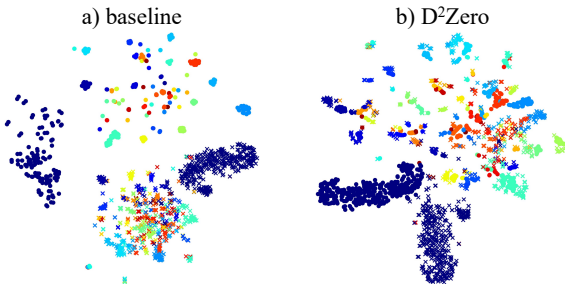


Figure 3. (Best viewed in color) t-SNE [50] visualization of image and text embeddings distribution on 48/17 split. The circle denotes the image embeddings. The cross denotes the text embeddings. Samples from different classes are marked in different colors.

We get the final results by combining all the components, which significantly surpass our baseline.

**Input-Conditional Classifier.** By replacing the conventional text embedding classifier with our input-conditional classifier, we can obtain significant improvement on unseen results, *e.g.*,  $\uparrow 4.8\%$  mAP on 48/17 split. The improvement on unseen group brings performance gain to HM results, *e.g.*, 6.91% HM-mAP and 3.68% HM-Recall gains on 48/17. Owing to our delicate design of classifier, the issues of bias toward seen categories and domain gap are greatly alleviated. Such significant improvements validate the superiority of our input-conditional classifier quantitatively.

In Figure 3, we utilize t-SNE [50] to visualize the image and text embeddings distribution with and without our input-conditional classifier. The t-SNE samples are from the same image with 4 classes. As shown in Figure 3(a), image embeddings (circles) and corresponding text embeddings (cross) are far away from each other, because of the domain gap between vision and language. In Figure 3(b), cross-modal features from same class, *e.g.*, the yellow circles and yellow cross, are pulled closer, showing high intra-class compactness and inter-class separability characteristics. Both quantitative and qualitative results demonstrate that with input-conditional classifier, image embeddings are well aligned with text embeddings and are capable of capturing discriminative features.

Method	Seen		Unseen		HM	
	mAP	Recall	mAP	Recall	mAP	Recall
word embedding bg	<b>53.49</b>	<b>77.52</b>	7.15	32.39	12.61	45.68
learned-parameter bg	53.01	76.21	8.74	34.41	15.01	47.41
<b>D<sup>2</sup>Zero</b> bg	53.17	76.13	<b>10.06</b>	<b>36.71</b>	<b>16.91</b>	<b>49.53</b>

Table 2. Comparison of different background (bg) designs.

Training Categories	AR@100		
	0.4	0.5	0.6
Seen	78.5	73.4	67.8
Seen + Unseen	82.1	78.4	73.1

Table 3. Instance proposal generalization ability.

Method	mAP	AP50	AP75
ZSI (w2v) [65]	0.008	0.009	0.008
<b>D<sup>2</sup>Zero</b> (w2v)	4.670	7.025	4.816
<b>D<sup>2</sup>Zero</b> (clip)	6.093	8.993	6.279

Table 4. Cross-dataset results on ADE20k validation dataset.

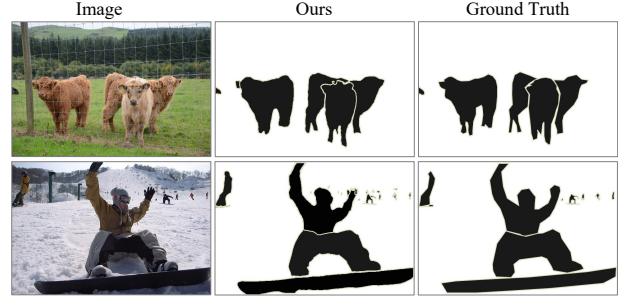


Figure 4. The predicted background masks by our approach can well exclude novel foreground objects of cow and snowboard.

**Image-Adaptive Background.** The different choices of background design have great impacts on the final performance, as shown in Table 2. The experiments are performed on our baseline and under 48/17 split. Learned-parameter bg surpasses word embedding bg with 2.40% HM-mAP and 1.73% HM-Recall, respectively, which indicates that learnable bg can mitigate background ambiguity to some extent. Compared with our image-adaptive background, using word embedding bg or learned-parameter bg, both mAP and Recall suffer from degradation, which verifies the effectiveness of our proposed approach.

In Figure 4, we visualize our generated background masks on unseen classes, *e.g.*, cow and snowboard. As shown in Figure 4, the foreground masks can be well segmented for both the seen and unseen classes. Because of the satisfactory generalization ability of mask proposal, we can generate a meaningful background mask for each image and produce a high-quality background representation.

**Unseen Cross-Entropy Loss.** When introducing pseudo unseen labels generated from the seen-unseen similarity of semantic embeddings, the performance is significantly improved by 6.05% HM-mAP and 5.12% HM-Recall over baseline (see Table 1  $\mathcal{L}^u$ ). This demonstrates that with the help of pseudo unseen labels, the feature extractor trained under the constraints of unseen categories can significantly alleviate bias towards seen classes issue and be well generalized to novel objects that have never show up in training.

**Generalization Ability of Instance Proposal.** In Table 3, we test the category-agnostic mask proposal of unseen classes at different IoU thresholds, using the model trained on “seen” and the model trained on “seen + unseen”. Training on “seen” achieves competitive results, demonstrating that Mask2Former can output masks for unseen categories when only trained with seen categories.

### 4.3. Transfer to Other Dataset

**D<sup>2</sup>Zero** model trained on COCO can be transferred to other instance segmentation datasets like ADE20k [66] via replacing semantic embeddings of our input-conditional classifier. We input the semantic embeddings of ADE20K classes as  $Q$  to our input-conditional classifier and testing

Split	Method (text encoder)	Seen		Unseen		HM	
		mAP	Recall	mAP	Recall	mAP	Recall
48/17	ZSI (w2v) [65]	43.04	64.48	3.65	44.90	6.73	52.94
	<b>D<sup>2</sup>Zero</b> (w2v)	52.53	75.66	9.48	37.93	16.06	50.52
	<b>D<sup>2</sup>Zero</b> (w2v-cp)	51.75	73.23	10.58	50.78	17.56	59.97
	<b>D<sup>2</sup>Zero</b> (clip)	<b>54.42</b>	<b>76.22</b>	15.06	38.38	23.59	51.06
	<b>D<sup>2</sup>Zero</b> (clip-cp)	54.12	73.22	<b>15.82</b>	<b>53.53</b>	<b>24.49</b>	<b>61.85</b>
65/15	ZSI (w2v) [65]	35.75	62.58	10.47	49.95	16.20	55.56
	<b>D<sup>2</sup>Zero</b> (w2v)	38.49	74.25	13.12	41.67	19.57	53.38
	<b>D<sup>2</sup>Zero</b> (w2v-cp)	37.32	70.43	15.39	58.64	21.79	63.99
	<b>D<sup>2</sup>Zero</b> (clip)	<b>41.18</b>	<b>74.94</b>	20.22	46.01	27.13	57.01
	<b>D<sup>2</sup>Zero</b> (clip-cp)	40.90	71.41	<b>21.91</b>	<b>65.72</b>	<b>28.54</b>	<b>68.45</b>

Table 5. Results on GZSIS. “cp” denotes copy-paste strategy of ZSI [65], *i.e.*, sharing instances between seen and unseen groups.

Split	Method (text encoder)	Recall@100			mAP
		0.4	0.5	0.6	
48/17	ZSI (w2v) [65]	50.3	44.9	38.7	9.0
	<b>D<sup>2</sup>Zero</b> (w2v)	60.0	55.9	50.8	16.1
	<b>D<sup>2</sup>Zero</b> (clip)	<b>65.5</b>	<b>61.4</b>	<b>55.9</b>	<b>21.7</b>
65/15	ZSI (w2v) [65]	55.8	50.0	42.9	10.5
	<b>D<sup>2</sup>Zero</b> (w2v)	68.5	65.1	60.6	16.9
	<b>D<sup>2</sup>Zero</b> (clip)	<b>73.3</b>	<b>69.7</b>	<b>64.9</b>	<b>23.7</b>

Table 6. Results on ZSIS.

our results on ADE20K, as shown in Table 4. Our method demonstrates good generalization ability on cross-dataset testing and significantly outperforms ZSI [65]. In ZSI [65], there are some classifier parameters related to the dataset category, making it impossible to transfer to other datasets.

### 4.4. Comparison with State-of-the-Art Methods

In Table 6 and Table 5, we follow the experimental settings in ZSI [65] to report our results on both the Zero-Shot Instance Segmentation (ZSIS) and Generalized Zero-Shot Instance Segmentation (GZSIS) tasks. The proposed **D<sup>2</sup>Zero** exceeds ZSI by a large margin, *e.g.*, our model with CLIP [46] as text encoder outperforms ZSI by 16.86% HM-mAP under the 48/17 split and 10.93% H-mAP under the 65/15 split. We also report our results using copy-paste strategy of ZSI, marked with “cp”. The “cp” strategy significantly improves recall performance for unseen classes but decreases precision since it brings many false positives. To further evaluate the superiority of our method, we conduct a model complexity comparison with ZSI. The #parameters/FLOPs of our **D<sup>2</sup>Zero** and ZSI [65] are 45.737M/227.7G and 69.6M/569.3G, **D<sup>2</sup>Zero** is dramati-



Figure 5. (Best viewed in color) From the 1st row to 3rd row are: ground truth, our results, and ZSI [65], respectively. ZSI fails to classify most of the unseen objects, e.g., cake in the first image and cow in the fifth image. And some novel objects are missed by ZSI due to background confusion, e.g., skateboard in the second image and couch in the third image. The proposed approach  $D^2Zero$  shows much better results by classification debiasing and background disambiguation.

Split	Method	Seen		Unseen		HM	
		mAP	Recall	mAP	Recall	mAP	Recall
48/17	DSES [2]	-	15.02	-	15.32	-	15.17
	PL [47]	35.92	38.24	4.12	26.32	7.39	31.18
	BLC [64]	42.10	57.56	4.50	46.39	8.20	51.37
	ZSI [65]	46.51	70.76	4.83	53.85	8.75	61.16
	$D^2Zero$ (w2v)	52.30	76.89	9.46	37.28	16.02	50.21
	$D^2Zero$ (w2v-cp)	51.31	72.04	10.55	54.14	17.50	62.01
	$D^2Zero$ (clip)	<b>54.47</b>	<b>77.52</b>	14.67	38.09	23.12	51.08
$D^2Zero$ (clip-cp)	54.14	74.09	<b>15.45</b>	<b>54.19</b>	<b>24.05</b>	<b>62.60</b>	
65/15	PL [47]	34.07	36.38	12.40	37.16	18.18	36.76
	BLC [64]	36.00	56.39	13.10	51.65	19.20	53.92
	ZSI [65]	38.68	67.11	13.60	58.93	20.13	62.76
	$D^2Zero$ (w2v)	38.71	74.25	13.00	41.41	19.46	53.16
	$D^2Zero$ (w2v-cp)	39.32	70.43	15.39	63.64	22.12	66.86
	$D^2Zero$ (clip)	<b>40.51</b>	<b>74.64</b>	20.23	46.33	26.99	57.17
	$D^2Zero$ (clip-cp)	40.23	70.98	<b>21.84</b>	<b>66.65</b>	<b>28.31</b>	<b>68.75</b>

Table 7. Results on GZSD. Previous methods all use word2vector.

cally more efficient than ZSI, thanks to our efficient mask proposal network and lightweight component design.

In Figure 5, we present a qualitative comparison with ZSI [65] for both seen and unseen classes on COCO under the 48/17 split. ZSI fails to classify most of the unseen objects. For example, ZSI identifies cow as horse. By contrast, our approach outputs more accurate instance labels for both seen and unseen categories and more precise mask predictions. Besides, our method successfully segments the objects missed by ZSI due to background ambiguity, like couch in the 3rd column, which demonstrates the effectiveness of our background disambiguation.

We also report our results on Zero-Shot Detection (ZSD) in Table 8 and Generalized Zero-Shot Detection (GZSD) in Table 7. We do not use bounding box regression but simply produce bounding box from our masks, and achieves new state-of-the-art performance on ZSD and GZSD. The above experiments and analysis all demonstrate

Split	Method	Recall@100			mAP
		0.4	0.5	0.6	
48/17	SB [2]	34.46	22.14	11.31	0.32
	DSES [2]	40.23	27.19	13.63	0.54
	TD [37]	45.50	34.30	18.10	-
	PL [47]	-	43.59	-	10.10
	Gtnet [62]	47.30	44.60	35.50	-
	DELO [67]	-	33.50	-	7.60
	BLC [64]	49.63	46.39	41.86	9.90
	ZSI [65]	57.40	53.90	48.30	11.40
	$D^2Zero$ (w2v)	60.00	56.10	52.00	16.30
	$D^2Zero$ (clip)	<b>65.70</b>	<b>61.70</b>	<b>57.70</b>	<b>21.40</b>
65/15	PL [47]	-	37.72	-	12.40
	BLC [64]	54.18	51.65	47.86	13.10
	ZSI [65]	61.90	58.90	54.40	13.60
	$D^2Zero$ (w2v)	69.10	66.20	62.30	16.80
	$D^2Zero$ (clip)	<b>73.90</b>	<b>70.70</b>	<b>66.60</b>	<b>23.50</b>

Table 8. Results on ZSD. Previous methods all use word2vector.

the effectiveness and efficiency of our  $D^2Zero$  on Zero-shot instance segmentation and detection tasks.

## 5. Conclusion

We propose  $D^2Zero$  with semantic-promoted debiasing and background disambiguation to address the two key challenges in zero-shot instance segmentation, i.e., bias issue and background ambiguity. To alleviate the bias issue, we introduce a semantic-constrained feature training strategy to utilize semantic knowledge of unseen classes and propose an input-conditional classifier to dynamically produce image-specific prototypes for classification. We discuss the background confusion and build an image-adaptive background prototype to better capture discriminative background clues. We achieve new state-of-the-art results on zero-shot instance segmentation and detection.

**Acknowledgement** This work was partially supported by National Natural Science Foundation of China (No.62173302).



## References

- [1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 2
- [2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 8
- [3] Supritam Bhattacharjee, Devraj Mandal, and Soma Biswas. Autoencoder based novelty detection for generalized zero shot learning. In *ICIP*. IEEE, 2019. 2
- [4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 1, 2
- [5] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *NeurIPS*, 32, 2019. 1, 2, 3
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*. Springer, 2020. 2
- [7] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Classifier and exemplar synthesis for zero-shot learning. *IJCV*, 128(1), 2020. 2
- [8] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*. Springer, 2016. 2, 3
- [9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 2, 3, 6
- [10] Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *ICCV*, October 2021. 3
- [11] Debasmrit Das and CS George Lee. Zero-shot image recognition using relational matching, adaptation and calibration. In *IJCNN*. IEEE, 2019. 2
- [12] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Iklizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *ICCV*, 2017. 2
- [13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 3
- [14] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE TPAMI*, 2022. 3
- [15] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 3
- [16] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 2
- [17] Rafael Felix, Ben Harwood, Michele Sasdelli, and Gustavo Carneiro. Generalised zero-shot learning with domain classification in a joint semantic and visual space. In *DICTA*. IEEE, 2019. 2
- [18] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 26, 2013. 2
- [19] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022. 3
- [20] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 6
- [21] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *ACM MM*, 2020. 1
- [22] Yuchen Guo, Guiguang Ding, Jungong Han, Xiaohan Ding, Sicheng Zhao, Zheng Wang, Chenggang Yan, and Qionghai Dai. Dual-view ranking with hardness assessment for zero-shot learning. In *AAAI*, volume 33, 2019. 2
- [23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6
- [25] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *CVPR*, 2023. 3
- [26] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. *NeurIPS*, 33, 2020. 3
- [27] Zhengdong Hu, Yifan Sun, and Yi Yang. Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [28] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*, pages 7020–7031, 2022. 3
- [29] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *ICLR*, 2017. 4
- [30] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, and Osamu Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*. IEEE, 2012. 2
- [31] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2):5453–5460, 2022. 2, 5
- [32] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*. IEEE, 2009. 2
- [33] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3), 2013. 2
- [34] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3

- [35] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *NeurIPS*, 33, 2020. 1, 2, 3
- [36] Yan Li, Zhen Jia, Junge Zhang, Kaiqi Huang, and Tieniu Tan. Deep semantic structural constraints for zero-shot learning. In *AAAI*, 2018. 2
- [37] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *AAAI*, 2019. 8
- [38] 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 *ECCV*. Springer, 2014. 6
- [39] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023. 3
- [40] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, pages 8741–8750, 2021. 2
- [41] Fengmao Lv, Haiyang Liu, Yichen Wang, Jiayi Zhao, and Guowu Yang. Learning unbiased zero-shot semantic segmentation networks via transductive transfer. *IEEE Sign. Process. Letters*, 27, 2020. 3
- [42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 5
- [43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 6
- [44] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *NeurIPS*, 22, 2009. 2
- [45] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation, 2021. 3
- [46] 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 *ICML*. PMLR, 2021. 5, 6, 7
- [47] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. *AAAI*, 2020. 8
- [48] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *TPAMI*, 35(7), 2012. 2
- [49] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *arXiv preprint arXiv:1301.3666*, 2013. 2
- [50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2
- [52] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*. Springer, 2020. 1, 2
- [53] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 3
- [54] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, 2019. 3
- [55] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41(9), 2018. 6
- [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] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 3
- [58] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013. 2
- [59] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022. 2, 3, 5
- [60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 6
- [61] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, 2021. 2, 3
- [62] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang. Gtnet: Generative transfer network for zero-shot object detection. *arXiv preprint arXiv:2001.06812*, 2020. 8
- [63] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *ACCV*, 2020. 2
- [64] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. *arXiv preprint arXiv:2010.04502*, 2020. 8
- [65] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7, 8
- [66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 7
- [67] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don’t even look once: Synthesizing features for zero-shot detection. In *CVPR*, 2020. 8