Cap2Seg: Inferring Semantic and Spatial Context from Captions for Zero-Shot Image Segmentation

Guiyu Tian[1], Shuai Wang[2], Jie Feng[2], Li Zhou[2], Yadong Mu[1]*
{wavey,myd}@pku.edu.cn,{wangshuai-hq,fengjie,zhouli_cto}@boe.com.cn
[1]Wangxuan Institute of Computer Technology, Peking University, [2]BOE Technology Group Co., Ltd.

ABSTRACT

Zero-shot image segmentation refers to the task of segmenting pixels from specific unseen semantic class. Previous methods mainly rely on historic segmentation tasks, such as using semantic embedding or word embedding of class names to infer a new segmentation model. In this work we describe Cap2Seg, a novel solution of zero-shot image segmentation that harnesses accompanying image captions for intelligently inferring spatial and semantic context for the zero-shot image segmentation task. As our main insight, image captions often implicitly entail the occurrence of a new class in an image and its most-confident spatial distribution. We define a *contextual entailment question* (CEQ) that tailors BERT-like text models. In specific, the proposed networks for inferring unseen classes consists of three branches (global / local / semi-global), which infer labels of unseen class from image level, image-stripe level or pixel level respectively. Comprehensive experiments and ablation studies are conducted on two image benchmarks, COCO-stuff and Pascal VOC. All clearly demonstrate the effectiveness of the proposed Cap2Seg, including a set of *hardest* unseen classes (*i.e.*, image captions do not literally contain the class names and direct matching for inference fails).

CCS CONCEPTS

• **Computing methodologies → Image segmentation**.

KEYWORDS

Neural networks, image segmentation, zero-shot learning

ACM Reference Format:
Guiyu Tian[1], Shuai Wang[2], Jie Feng[2], Li Zhou[2], Yadong Mu[1]. 2020. Cap2Seg: Inferring Semantic and Spatial Context from Captions for Zero-Shot Image Segmentation. In *28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413990

1 INTRODUCTION

Zero-shot image segmentation is a recently-emerging task in computer vision and multimedia analysis. It aims to develop models to segment novel objects with no labeled training data. This task is derived from zero shot learning task, yet with a special emphasis of spatial correlation of different image pixels. Only a very limited number of related works have been devoted to this novel task, as can be found in [6, 21, 49]. Among them, the mainstream approach employs similar idea in general zero-shot learning. Class names are embedded into a common subspace according to semantic affinities, where a novel class can be linearly represented by a few known ones. This paves the way of transferring knowledge from known classes to unknown class.

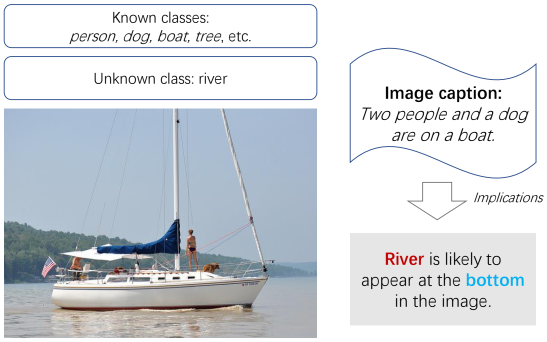
Known classes:
person, dog, boat, tree, etc.
Unknown class: river
Image caption:
Two people and a dog are on a boat.
Implications
River is likely to appear at the **bottom** in the image.

Figure 1: Illustration of the motivating fact of Cap2Seg. Our key observation is that image captions may implicitly convey the occurrence of a novel semantic class (highlighted in red text) and spatial location (in blue text).

This work addresses zero-shot image segmentation from a rarely-explored aspect. To combat the scarcity of pixel-level annotations, we propose to explore the accompanying image captions for inferring semantic / spatial context for unseen classes. The noisy annotations harvested from image captions for unseen classes are then fed into a segmentation model. In general, the salient objects in the image may be literally mentioned in the caption. For instance, if the caption is "a giraffe is drinking water", we can faithfully expect a scenario with some instances of giraffes. An image caption tends to concisely refer to only a few eye-catching objects that skeleton the scene, leaving other interested objects undetectable via direct word matching in captions. Nonetheless, one can intuitively infer novel classes via semantic common sense. For example, the above-mentioned "a giraffe is drinking water" implies high likelihood for *tree* or *grass*, in comparison with *television* or *desk*.

Furthermore, the spatial layout of known classes in an image also provides clues for the position of some unknown class. As seen in Figure 1, given a number of spatially-annotated concept such as person, dog, etc., we can infer that the instance of river is likely to appear around the boat, or state differently in the bottom part

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA.

ACM ISBN 978-1-4503-7988-5/20/10...$15.00
https://doi.org/10.1145/3394171.3413990

of the image. This example illustrates captions and knowledge of spatial layout of known classes can collaboratively infer whether and where unknown classes would appear in an image.

Based on the observation described above, we propose Cap2Seg, a novel inference engine that can harvest useful albeit noisy annotations for an unseen semantic class. It is implemented by a three-branch network which reads image captions and confidence scores of a few known classes. A global branch directly infers from captions an image-level likelihood that current image contains the interested unseen class. A local branch is devised to transfer the knowledge of known classes and provide a rough estimation of unseen class at the pixel level. A semi-global branch aims to provide spatial inference, which splits an image into even horizontal stripes and estimate the probability of unseen class over the stripes. After the inference is accomplished, since the inferred semantic / spatial context are highly noisy, we develop a robust segmentation model, which is jointly trained on all classes and utilize spatially-weighted pair-wise ranking loss to resist annotation noise. To our best knowledge, it is the first work that image captions are explored in the task of zero-shot segmentation. We strongly believe that this brings fresh air to existing semantic embedding based methods. Our comprehensive evaluations on both COCO-stuff and PASCAL VOC clearly re-calibrate new state-of-the-art performances for the task of zero-shot image segmentation.

2 RELATED WORKS

Zero shot classification and segmentation. Zero shot learning aims to enable model to recognize objects which never appear during training time [18, 45, 50]. The mainstream method usually utilize predefined semantic embedding (*i.e.*, class attributes [26] or word embedding[20, 32]) to transfer the learned knowledge from known classes to unknown classes. Generalized zero shot learning has also been investigated to prevent model from forgetting the knowledge already learner from seen classes [9, 30, 42]. Most of the works focus on zero shot image recognition. Recently, several works has expand zero shot learning from image classification to segmentation [6, 49].

Weakly-supervised semantic segmentation. Modern semantic segmentation framework [3, 10, 31] is built on fully convolution neural network. However, collecting a large number of pixel-level annotations for training semantic segmentation models is time consuming and laborious. In order to reduce reliance on dense annotations, many works have begun to investigate models trained from weaker supervised signals, including bounding boxs [14, 34], scribbles [28, 51], points [4], etc. In addition, our work is more related to image-level supervised methods [24, 25, 35, 38, 46, 57], where segmentation model typically learns with object tags paired with image. Existing approaches usually use the class activation map [56] to obtain the most discriminative class response as the label region, which can be refined to make the supervision stronger [25, 46, 57].

Caption-guided visual learning. There are also some works which use auxiliary text material to augment vision learning. TAM-NET [40] utilizes text to generate text activation maps, which can be used for augmenting class activation map in segmentation task. Cap2Det [52] leverages the signal that captions provide for weakly supervised detection. However, caption-enhanced image segmentation models are still inadequately explored in the literature.

3 OUR PROPOSED MODEL

3.1 Task formulation

Following previous practice in [6, 21, 49], let us use the notations $\mathcal{S}, \mathcal{U}$ for the collections of seen / unseen class names, respectively. $\mathcal{S} \cap \mathcal{U} = \emptyset$. Let $\mathcal{D} = \{(x, y, t)\}$ be the set of data. For each tuple, x indexes an image from some image corpus $\mathcal{I}$. y represents the label mask provided in a pixel-wise manner, with only seen classes visible during training and all annotations for $\mathcal{S} \cup \mathcal{U}$ revealed on final model evaluation. $t = caption(x)$ is a free-form textual caption for image x, available for both training and testing images. The ultimate goal of zero-shot image segmentation is to learn a model from seen classes which can generalize to pixel-level prediction on arbitrary unseen classes. Essentially, our setting is identical to that used in ZS3Net [6] or SPNet [49]. Training images may contain pixels of unseen classes, keeping their annotations anonymous. Nonetheless, self-supervision was further mobilized in [6] to include additional cues. In particular, the trained ZS3Net was used for scanning training images, and the top p% of the most confident among detected pixels for unseen classes provide new training features, leading the so-called ZS5Net. For fair comparison, we discard such post-enhancement in all experiments.

There are two popularly-adopted evaluation protocols in previous development of zero-shot image segmentation [6], primarily differing in the label space (either $\mathcal{U}$ or $\mathcal{S} \cup \mathcal{U}$) when processing a novel image. The latter case predicts both seen / unseen classes and is termed as *generalized zero-shot image segmentation* in the literature. In our evaluations both protocols will be considered.

3.2 Model Overview

Fig. 2 depicts the architecture of our proposed Cap2Seg model. This unified model is comprised of two crucial components: an inference machine that is capable of reading in an image accompanied with textual caption and predicting image-specific visual occurrence / spatial distribution of unseen classes, and a detector which is trained from caption-induced weak supervision using spatially-weighed pairwise ranking loss.

3.3 Inference From Captions

As exposed in Fig. 1, one of the key missions in Cap2Seg is to squeeze the likelihood of an image's containing pixels from specific unseen classes and the most confident image regions where these pixels reside. The later section separately details our proposed solution for this mission.

3.3.1 Visual Occurrence Estimation By Contextual Entailment. Let us first briefly summarize the manipulation of image captions in most relevant works Cap2Det [52] or TAM [40] for image segmentation or detection. The input to caption processor is obtained by encoding each word with a word2vec model and average pooling over words, often intertwined with fully-connected layers for fine-tuning. Both suffer from two key weaknesses: first, they are intrinsically designed for a fixed-set problem. For example, the set of all interested pseudo-labels are pre-specified before the model

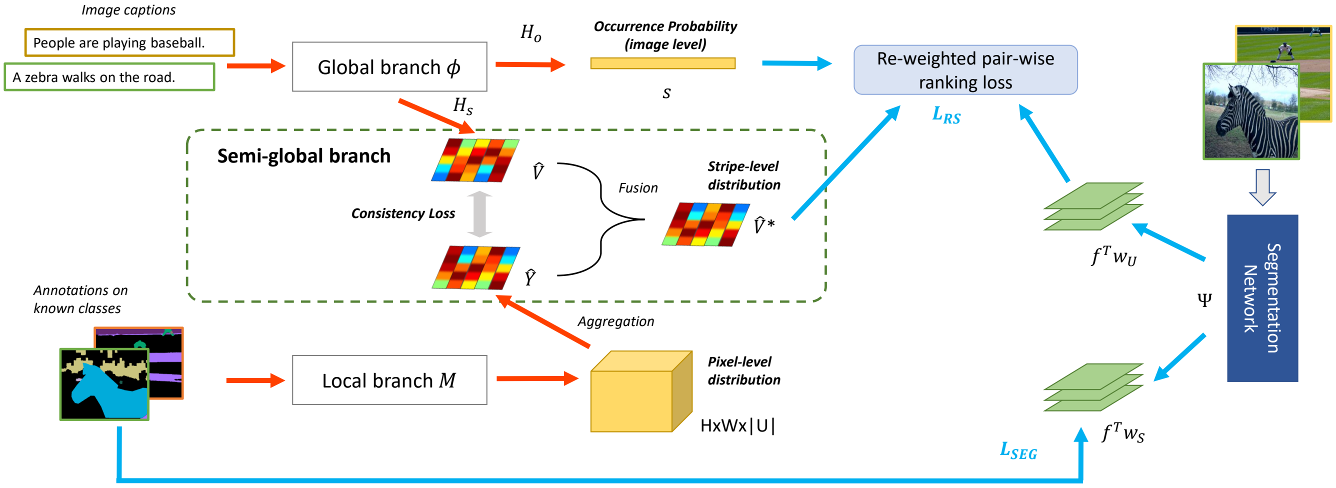

Figure 2: Architectural design of of our proposed Cap2Seg inference engine (in Sections 3.3-3.5) and follow-up image segmentation model (in Section 3.6).

optimization starts. TAM [40] relies on hand-crafted *compound concepts* extracted from known captions. Neither can be trivially generalized to arbitrary unseen class. Secondly, the simple pooling step of all word embeddings arguably discard the contextual information, which is crucial for inferring some not explicitly mentioned novel class.

Motivated by above observations, we opt for the pre-trained language model, BERT [49], as the major workhorse of our caption-inferring model. Our preference of BERT over other alternatives is its utilization of bidirectional training of transformers, which gives it deeper sense of language context and flow. When BERT is fed with an image caption t, it encodes the entire sentence into a single feature vector, hereafter denoted by $\phi(t)$. This context-preserving property makes our model divergent from related caption-reading methods [40, 52].

Our second technical novelty related to caption inference is *contextual entailment*, which is designed for answer such a question: *if an image caption is known to be true, is the image also highly likely to contain some novel class?* Our solution of inspired by the Natural Language Inference (NLI) or Recognizing Textual Entailment (RTE) [13] in natural language processing. It refers to the task of classifying a pair of *premise* and *hypothesis* sentences into three classes: contradiction, neutral, and entailment. For example, "a soccer game with multiple males playing" entails "some men are playing a sport", and contradicts "no men are moving in the image".

In the context of zero-shot image segmentation, above entailment is defined on each individual image x. Let $caption(x)$ be the associated caption sentence and $description(c)$ be the textual representation of some class c. We regard the caption as the premise, and the class description as hypothesis, obtaining the following *contextual entailment question* (CEQ):

$$\mathbf{CEQ}(x, c) \quad : \quad caption(x);\ [EOS];\ description(c), \tag{1}$$

where $[EOS]$ is a token that implies the end of a sentence. This way, the task of estimating pixel occurrence of some novel class c in an image is transformed to a more elegant, tractable form that fully harnesses the contextual modeling of BERT. The goal boils down to predicting the relation between premise and hypothesis, either entailment or contradiction. A judgement of highly-confident entailment implies that class c is consistent with the image caption's semantics. Similar argument for a judgement of contradiction. In particular, we relax the answer to CEQ to be softly between $[0, 1]$, casting it to be a confidence-modulated binary classification.

This can be empirically accomplished by appending a fully-connected head (denoted as $H_o(\cdot)$ on top of the backbone of BERT model. Let $s_{x,c}$ be the probability that some pixels of class c appear in image x, which is calculated as:

$$s_{x,c} = sigmoid\left(H_o(\phi(caption(x);\ [EOS];\ description(c)))\right). \tag{2}$$

Critically, the optimization of head H_o and fine-tuning ϕ are conducted over known classes $\mathcal{S}$. We adopt a loss of binary cross entropy as below:

$$\mathcal{L}_o = \sum_{x \in \mathcal{I}} \sum_{c \in \mathcal{S}} -[I(c \in y(x)))\cdot log(s_{x,c}) + (1 - I(c \in y(x)))\cdot log(1 - s_{x,c})], \tag{3}$$

where $y(x)$ is the label mask for image x. The indicator function $I(c \in y(x)))$ returns 1 if class c occurs in this specific image, otherwise 0.

During deploying the learned model, we can construct a CEQ by pairing a captioned image with some unknown class $c_u \in \mathcal{U}$, simply by specifying the class-related $description(c_u)$ (*e.g.*, class names) in the hypothesis part. Compared with traditional method of co-embedding visual / word vectors, our proposed CEQ fully takes advantage of context-rich caption and has superior flexibility in tackling novel class inquiry.

3.3.2 Spatial Distribution Estimation. The spatial arrangement of different objects is crucial in image segmentation. There are at least two cues that we can utilize for guessing the position of some object in an image. The first cue is inter-object structural arrangement. For instance, it is common to observe 'person' in front of 'desk' but rarely see a 'giraffe' doing so. Secondly, some objects or concepts tend to have concentrated spatial distribution, like that 'sky' is more frequently seen on the top region of an image. In conventional image segmentation model such as DeepLab [10], such spatial priors can be effectively learned and enforced via techniques like conditional random fields.

In this work we take an unexplored step of inferring image-specific class-wise spatial distribution from image captions. The assumption is that, complex context in the captions can roughly tell where an object is located. Implementation of this idea is still based on the customization of BERT model. A trivial observation is, in most cases an image and its horizontally-mirrored version can be described by a same caption. This is supposed to significantly complicate the prediction of horizontal position of an object. Therefore, our model only focuses on vertically localizing some object in an image. In particular, all images will be split into K horizontal stripes of uniform length. For an image x, let $g^{(k)}_{x,c}$ be the number of image pixels in the k-th stripe that is categorized to be class c. This forms a distribution for class c over all stripes, namely

$$V_{x,c} = \left\{ q^{(k)}_{x,c} \mid k = 1 \ldots K \right\} = \left\{ g^{(k)}_{x,c} / \sum_k g^{(k)}_{x,c} \mid k = 1 \ldots K \right\}, \quad (4)$$

where $q^{(k)}_{x,c}$ is normalized $g^{(k)}_{x,c}$.

We simply append another head $H_s(\cdot)$ to the backbone of BERT model, which is devised for estimating the spatial distribution of some class c in an image x. The estimation can be obtained as

$$\begin{aligned} \widehat{V}_{x,c} &= \left\{ \hat{q}^{(k)}_{x,c} \mid k = 1 \ldots K \right\} \\ &= softmax(H_s(\phi(caption(x);\ [EOS];\ description(c)))). \end{aligned}$$

During training, $H_s(\cdot)$ is iteratively optimized by minimizing the distributional discrepancy between all pairs of $V_{x,c}, \widehat{V}_{x,c}$ over all seen classes in $\mathcal{S}$. This is achieved via an informational entropy objective:

$$\mathcal{L}_s = \sum_{x \in \mathcal{I}} \sum_{c \in \mathcal{S}} \sum_{k=1 \ldots K} -\hat{q}^{(k)}_{x,c} \log(q^{(k)}_{x,c}). \quad (5)$$

Overall, we purse both image-specific class-wise visual occurrence and spatial distribution by fine-tuning the BERT model, with a unified optimization target:

$$\mathcal{L} = \mathcal{L}_o + \mathcal{L}_s, \quad (6)$$

where H_o, H_s, controlled by $\mathcal{L}_o + \mathcal{L}_s$, prove empirically complementary to each other, as corroborated later in our experiments.

3.4 Local Pixel Level Inference

The global contextual entailment model described above utilizes the global semantic context of the caption annotation to infer the occurrence probability and vertical spatial distribution of unknown classes. In order to make full use of the pixel wise annotation of the known classes in the image, we design a local pixel-level prediction network, starting from a single source pixel of the known class and predicting the class information of the target pixel. Finally, the prediction results from all of the known classes pixel will be integrated for the spatial distribution of unknown classes.

Given an image x, our local pixel level inference model M samples a source pixel s of known class where $c_s \in \mathcal{S}$ from all the labeled pixels x_l, and an unlabeled target pixel t. The probability that $c_t = u \in \mathcal{U}$ is inferred by M as:

$$P(c_t = u \mid x, c_s) = sigmoid(M(pos(s), pos(t), w_{c_s}, w_u)), \quad (7)$$

where $pos(p)$ is the 2D coordinates of pixel p resized to $[0, 1]$. $w_c \in \mathbb{R}^d$ is the word embedding related to class c. In practice, M is implemented as a multi-layer perceptron (MLP) network, where we stack the locations and word embeddings as a 1D input vector.

The spatial distribution of unknown class u can be integrated by the prediction results from all the labeled pixel which includes a variety of class categories and locations:

$$P(c_t = u \mid x) = \sum_{s \in X_l} sigmoid(M(pos(s), pos(t), w_{c_s}, w_u)) \ / \ |X_l|, \quad (8)$$

where X_l denotes the set of labeled pixels from known classes and $|X_l|$ is set carnality used as a rescaling factor. By this means, the spatial distribution of certain unknown class u can be obtained by the pixel-wise annotations of known classes. For reducing the computational complexity, we also conduct sub-sampling on the labeled pixels, where we only look into $n_{sub} << |X_l|$ labeled pixels for each image. In the experiments, we found even a small n_{sub} ,e.g. 20, can achieve a good performance.

Training $M(\cdot)$ does not require any annotation of unknown class. In each iteration, we randomly choose a known class u' to temporarily play the role of an unknown class, and predict its distribution based on other known classes:

$$\begin{aligned} \mathcal{L}_p = -\sum_{x \in \mathcal{I}} \sum_{t \in X_l} \sum_{u' \in \mathcal{S}} & \quad I(c_t = u') log(P(c_t = u' \mid x)) \\ & + \quad (1 - I(c_t = u')) log(1 - P(c_t = u' \mid x)). \quad (9) \end{aligned}$$

3.5 Semi-Global Branch with Consistency Loss

The global contextual entailment takes caption annotation as input which contains the global semantic information, while the local pixel-level inference model take pixel-wise annotation of known classes as input which contains local class information. In order to make the two kind of information in different modalities and different scales complementary to each other, we develop a semi-global branch with a consistency loss to fuse the inferred spatial distribution from two branch.

The consistency loss is built on the horizontal stripes from the contextual entailment and the pixel-wise spatial distribution predicted by pixel-level inference model $M(\cdot)$. For a certain class c, based on the spatial distribution given by $M(\cdot)$, we average the pixels in each horizontal strip to obtain the vertical distribution $\widehat{Y}_{x,c}$:

$$\begin{aligned} \widehat{Y}_{x,c} &= \left\{ \hat{y}^{(k)}_{x,c} \mid k = 1 \ldots K \right\} \\ &= softmax \left\{ \sum_{t \in strip(k)} P(c_t = c \mid x) \mid k = 1 \ldots K \right\}. \quad (10) \end{aligned}$$

Then we take the L2 distance between these two distribution as the objective function:

$$\mathcal{L}_{consist} = \sum_{x \in \mathcal{I}} \sum_{c \in \mathcal{S}} ||\widehat{Y}_{x,c} - \widehat{V}_{x,c}||_2. \quad (11)$$

During deploying the learned model, the final predicted spatial distribution of unknown class can be obtained by fusing these two complementary distributions, namely $\widehat{V}^*_{x,u} \leftarrow (\widehat{Y}_{x,u} + \widehat{V}_{x,u})/2$.. The proposed Cap2Seg inference engine is obtained by jointly optimizing all objectives from three branches in Figure 2, including the losses in Equations (6)(9)(11).

3.6 Learning the Segmentation Model for Unseen Classes

In this section, we describe a deep model that uses the harvested weak supervision ($s_{x,c}$ for visual occurrence and $\widehat{V}^*_{x,c}$ for stripe-level spatial distribution) for the generalization to unseen classes.

3.6.1 *Pixel-wise Segmentation Loss on Seen Classes.*

The first loss is defined on seen classes and utilizes purely visual information. A majority of modern image segmentation methods adopt an encoder-decoder neural architecture, which is adopted in this work. For ease of statement, we re-interpret such architecture in a “feature extractor + linear classifier" point of view. Specifically, given an image $x \in \mathcal{X}$, assume the penultimate layer of a typical such network generates a feature map $f = \psi(x)$, where $\psi(\cdot)$ encapsulates all involved neural operations. Let $f \in \mathbb{R}^{h \times w \times d}$, where $h \times w$ defines the spatial resolution and d is the extracted feature's length. The prediction in the image segmentation task is performed in a pixel-wise fashion. Following the practice in [49], we use $w_c \in \mathbb{R}^d$ be a fixed word2vec embedding related to class $c \in \mathcal{S} \cup \mathcal{U}$. For a pixel (i, j) in the feature map, we adopt a standard softmax function as below:

$$P(\hat{y}_{i,j} = c) = \exp(f^T_{i,j} w_c) \ / \sum_{c \in \mathcal{S}} exp(f^T_{i,j} w_c), \tag{12}$$

where $\hat{y}_{i,j}$ is the predicted label of pixel (i, j) in image x (whenever causing no confusion we omit the index of x).

During the training time, the segmentation model is trained only on pixels from the known class set $\mathcal{S}$ with cross entropy loss:

$$\mathcal{L}_{seg} = -\sum_{x \in \mathcal{I}} \sum_{i,j} I[y_{i,j} \in \mathcal{S}] \cdot \log(P(\hat{y}_{i,j} = y_{i,j})), \tag{13}$$

where the indicator function $I[y_{i,j} \in \mathcal{S}]$ outputs 1 if the corresponding pixel is annotated to some seen class.

For an unseen class $c \in \mathcal{U}$, the corresponding discriminating vector w_c is also pre-fixed via word2vec. We allow only $\psi(\cdot)$ is learnable via gradient back-propagation. In a typical setting of zero-shot image segmentation, a pixel (i, j) in a testing image during the inference stage is calculated via:

$$\hat{y}_{i,j} = \arg\max_{c \in \mathcal{U}} f^T_{i,j} w_c, \ or \ \hat{y}_{i,j} = \arg\max_{c \in \mathcal{S} \cup \mathcal{U}} f^T_{i,j} w_c, \tag{14}$$

where the latter is for generalized zero-shot segmentation.

3.6.2 *Spatially-Weighed Ranking Loss on Unseen Classes.*

Given only the caption annotation t of an image x, we can obtain the occurrence probability $s_{x,c}$ that class c would appear in image x. We propose a pair-wise ranking loss to utilize $s_{x,c}$. For a training image x, we have the access to the ground truth label map y on known classes S. Let Y be the set of unlabeled pixel positions:

$$Y = \{(i, j) \mid y_{i,j} \notin S\}. \tag{15}$$

Given a pair of image x_1,x_2, the encoded feature map f_1, f_2 can be obtained with the CNN model $\psi(\cdot)$. Occurrence probabilities of specific class $s_{1,c}$, $s_{2,c}$ can be inferred with our inference model using the caption annotation t_1, t_2. $s_{1,c_u} > s_{2,c_u}$ implies that image x_1 is more likely to contain class c_u than image x_2. Or equivalently, the unlabeled part Y_1 of x_1 is more likely to contain unknown class c_u than the unlabeled part Y_2 of x_2. The proposed ranking loss can be written as:

$$\begin{aligned} L_R = -\sum_{c_u \in U} I(s_{1,c_u}, s_{2,c_u}) (\frac{1}{|Y_1|} \sum_{(i,j) \in Y_1} f_1(i, j)^T W^u_{c_u} \\ - \frac{1}{|Y_2|} \sum_{(i,j) \in Y_2} f_2(i, j)^T W^u_{c_u}), \end{aligned} \tag{16}$$

where $I(s_1, s_2)$ is the indicator function. $I(s_1, s_2) = 1$ if $s_1 > s_2$ otherwise -1.

The horizontal stripe-level distribution of any unknown class can be inferred from the captions as described in Section 3.5. Intuitively, such information can be used to re-weigh the stripe-like image regions. L_R can be further improved to a weighted version:

$$\begin{aligned} L_{RS} = -\sum_{c_u \in U} I(s_{1,c_u}, s_{2,c_u}) (\sum_{k \in (1,2,\ldots,N)} \frac{{p_1}^{c_u}_k}{|Y_{1,k}|} \sum_{(i,j) \in Y_{1,k}} f_1(i, j)^T W^u_{c_u} \\ - \sum_{k \in (1,2,\ldots,N)} \frac{{p_2}^{c_u}_k}{|Y_{2,k}|} \sum_{(i,j) \in Y_{2,k}} f_2(i, j)^T W^u_{c_u}), \end{aligned} \tag{17}$$

where $k \in (1, 2, ..., K)$ is the index of horizontal strips. $\{p^{c_u}_k | k \in (1, 2, ..., K)\}$ is the predicted spatial distribution $\widehat{V}^*_{x,c_u}$ of class c_u. For simplicity of the formula, here we omit the traversal of the paired images x_1, x_2.

3.6.3 *Overall Loss.*

During the training stage, the segmentation model is trained with loss:

$$L = L_{SEG} + \lambda L_{RS}, \tag{18}$$

where λ is the weight factor to balance the loss L_{SEG} from known class and L_{RS} from unknown class.

4 EXPERIMENTS

4.1 Dataset

We conduct our experiments on two segmentation datasets with both captions and pixel labels: **COCO-stuff** [7] has 164K images from the popular COCO dataset with pixel-wise annotations among 182 classes. It also provides 5 sentences per image. For the unknown class split of COCO-stuff, a class split is proposed in [49] with 167 known classes and 15 unknown classes, yet most of the unknown classes are categories often directly mentioned by captions. To study the capability of inferring implicit semantics in image captions, we rank all classes according to the frequencies of being directly mentioned by caption. 15 classes are chosen as unknown classes, covering low, medium and high explicitly-mentioned cases. The selected unknown classes and the probability of their class name's being directly mentioned in captions are shown in Figure 3. Importantly, ImageNet [39] pretrained network is leveraged to initialize the segmentation model. For preventing label leaking, the 15 unknown classes have no intersection with the 1000 classes in ImageNet. **Pascal VOC** [16] has 12k images with 20 object categories. There are 906 training images, which evenly distributes over

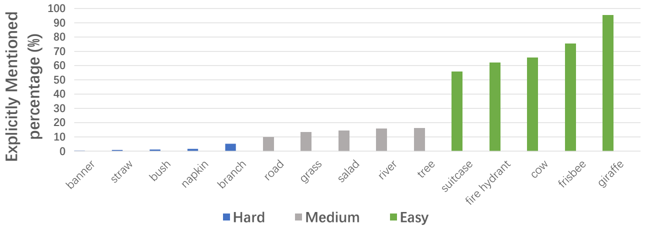

Figure 3: On COCO-stuff, all classes are ranked according to the probability of being directly mentioned by captions. We select three distinct subsets at different difficulty levels as the unknown classes.

all 20 categories, are equipped with caption annotations [36]. In practice, we first train a segmentation model on the rest of training images on the known classes, and then fine-tuning the model on the 906 images with occurrence probability and spatial distribution of unknown classes estimated from our inference model.

4.2 Implementation details

We adopt widely-used BERT [47] as the backbone of the inference model. Adam [22] solver is used, with the learning rate initialized as 3e-5. Regarding the segmentation model, we follow the settings in [49]. In specific, DeepLab-v2 [11] with ResNet backbone is used as the basic model. Standard SGD (momentum: 0.9, weight decacy rate: 0.0005, and learning rate is initialized as 1.5e-4 with a poly decaying strategy) is adopted. n_{sub} is set as 20 to make a trade off between performance and computational complexity. We split 5K images from training set as validation set to choose the hyperparameter λ in Equation (18). λ is empirically fixed to 0.01 in the experiment.

For occurrence probability inference, we use *mean average precision* (mAP) as the metric. For spatial distribution inference, we use cosine similarity between normalized predicted distribution and groundtruth as the metric. For the experiments on segmentation, *mean intersection over union* (mIoU) among unknown classes and our three subsets is chosen as the metrics in zero shot learning setting (ZSL). In generalized zero shot learning setting (GZSL), as suggested in [49], we report the mIoU on known classes, unknown classes and the harmonic mean of them:

$$H = \frac{2 * mIoU_{known} * mIoU_{unknown}}{mIoU_{known} + mIoU_{unknown}}. \quad (19)$$

5 EVALUATION

5.1 Occurrence Probability Inference

For evaluating the proposed contextual entailment model, we implement several alternative strategies:

Exact match: This method gives a binary occurrence result by directly matching the class name and its variants, *e.g.*, synonyms and plurality, with every noun in the caption. Thus, the occurrence probability is 1 if the match successes, otherwise 0. As discussed before, this method can only infer the classes that are easily captured directly by caption.

Sentence pooling: This method encodes the caption into a feature vector by pooling the word embedding of every word in the caption. Then the occurrence probability is defined as the cosine similarity between encoded feature and the word embedding of certain class. This method can capture the semantic information of caption in a coarse-grained manner.

Method	Hard	Medium	Easy	Average
Exact Match	5.4	33.5	72.8	37.2
Sentence Pooling	7.7	37.1	65.3	36.7
Traditional ZSL [37]	8.6	52.0	**86.6**	49.0
Context Entailment	**14.2**	**52.3**	85.6	**50.7**
CE with Multi-scale	15.1	52.7	85.7	51.2
CE with Multi-scale and curriculum	**15.8**	**52.9**	**87.0**	**51.9**

Table 1: Occurrence probability estimation of different inference method in terms of mAP on COCO-stuff.

Method	Hard	Meidum	Easy	Average
SPNet[49]	20.69	60.09	26.81	35.86
Exact Match	21.56	61.10	27.76	36.80
Traditional ZSL[37]	21.57	61.70	28.18	37.14
Context Entailment	**22.96**	**62.43**	**28.35**	**37.91**

Table 2: Zero-shot image segmentation results obtained from the occurrence probability by different inference methods in terms of mAP on COCO-stuff. Note that here image segmentation models are trained without stripe-level estimation, namely L_R loss is used.

Traditional ZSL: Since the problem we are facing is essentially a zero shot learning problem, where the samples of unknown classes is absent during training, we also adopt the traditional ZSL method as one of our baselines. This method firstly uses language model (such as BERT) to encode caption into a vector, and treats the cosine similarity between this vector and the word embedding of interested class as the probability of occurrence. The model to encode the caption is trained over known classes $\mathcal{S}$ and performs prediction on unknown classes $\mathcal{U}$.

The occurrence probability inference result is shown in Table 1. The directly-mentioned percentage has a great impact on the performances. As clearly seen, mAP scores on the easy set is far higher than the hard set. Even the naive method *Exact Match* can achieve striking results on the easy set (72.8%). For *Exact Match* and *Sentence Pooling* that require no training, inference performances on the hard set are notably inferior (5.4% and 7.7%, respectively), indicating their limitation in tackling classes not mentioned in image captions. The *Traditional ZSL* method has a slightly stronger inference performance on the hard set than the first two methods (8.6%). In contrast, with the contextual entailment (denoted as CE in Table 1) technique, our proposed model's inference ability on the hard set has been significantly improved in comparison with all baselines (14.2%). We attribute the success to better integration of the caption-induced global semantic information and the word embedding representation of the classes.

Inspired by [29], we find that the fusion of multi-scale features of the BERT-based language model can make further improvement to the inference performance. BERT is known to be a multi-layer transformer. To get a more powerful feature from BERT, we average the low-level features, which is sensitive to the contained text words, with high-level features, which is sensitive to the whole semantic context. Experiments in 1 show the clear accuracy promotion brought by multi-scale feature.

Another source of improvement stems from curriculum learning [5]. Curriculum learning refers to a training policy that starts

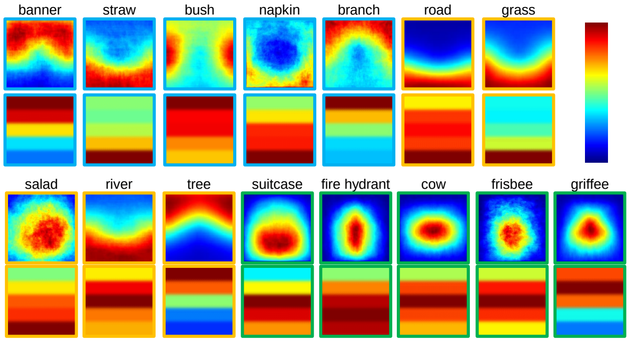

Figure 4: Visualization of inference results of the semi-global branch in Cap2Seg and ground truth in the form of heat-maps. For each class, the heat-map on the first row shows the ground truth aggregated from all images in this class. The second row represents the averaged estimation returned by our semi-global inference engine. Best viewing in the color mode.

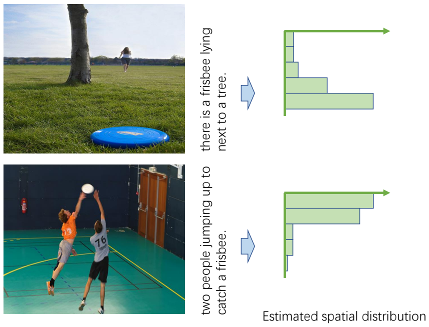
there is a frisbee lying next to a tree.
two people jumping up to catch a frisbee.
Estimated spatial distribution

Figure 5: Visualization of stripe-level spatial distribution estimation. For two images from the same class *frisbee*, the semi-global inference branch will return distinct estimations according to different captions.

with simple samples and then gradually increases the sample difficulty. This kind of training policy has demonstrated to be beneficial when the learning difficult of different samples varies a lot, which fits our case. During training, we first limit the active samples to be from easy classes, then gradually add samples from medium and hard classes into the training set. Experiments in 1 also show the improvement brought by this training policy.

The impact of different inference methods on the zero-shot image segmentation is shown in Table 2. As observed, the final segmentation performance highly hinges on the quality of inferred occurrence probability. Our proposed contextual entailment method significantly surpasses all the baselines.

5.2 Spatial Distribution Inference

Withe the semi-global branch, the Cap2Seg inference engine can predict the spatial distribution on horizontal stripes for an unknown class, mainly using the high-level semantic information in caption and word embedding of the class name. Visualization of such estimation on unknown classes are shown in Figure 4, where the class-wise heat-map is aggregated over all images from this class.

	Global	Local	Semi-Global	Local (AP)
w/o consistency loss	0.525	0.486	0.534	14.0
w/ consistency loss	0.538	0.498	**0.543**	16.0

Table 3: Comparisons of different training objectives and the results of fusing two branches on the spatial distribution prediction on COCO-stuff. In most cases the reported scores are Cosine similarity between estimated and ground-truth stripe-level distributional vectors. Higher scores are favored. However, note that the results of the local branch is pixel-wise, thus average precision (AP) serves the metric.

As seen, the distributions of many classes are consistent to human cognition. For instance, *banner* and *tree* are likely to occur on the top of the image. *road* and *grass* tend to appear on the bottom of the image. Although only trained among known classes, our proposed semi-global branch proves to effectively capture the spatial distribution of different unknown classes. In addition, we would also emphasize that the spatial estimatin is image-adaptive. Even for images from the same class, the estimations returned by the semi-global branch may differ significantly under different semantic context. Some concrete examples are presented in Figure 5.

We also discuss the selection on region number K. In order to align the results of different region numbers, we use the improvement of mean cosine similarity of the predicted result relative to the random result as the metrics. The improvement results are 0.13, 0.15, 0.19, 0.17, when the region number is 3, 4, 5, 6, respectively. The improvement is relatively small with a small region number, i.e., 0.13 for 3, because of the coarse-grained division. The improvement will also decreases when the region number is too large, due to the superposition of prediction noise. Thus, we use 5 as the region number in all of the following experiments.

Our local branch, a pixel-level inference model, takes the pixel-wise annotation of known class as input, and predicts the occurrence probability of unknown class on each background pixel. The pixel-level occurrence probability of certain unknown class can be assembled into the same format as the vertical distribution predicted by global inference model. During the training stage, we add a consistency loss between these two distributions to encourage a consistent result. Since these two kind of spatial distribution are inferred from the information of different modalities, our semi-global branch fuses these two kind of spatial distributions to obtain the final supervision signal for the segmentation model. Experiments are conducted to evaluate the performance of spatial distribution inference of our model and the effect of proposed consistency loss. Cosine similarity is adopted between predicted vertical distribution and the ground truth as the metric to evaluate the performance of vertical distribution prediction. For the local pixel-level inference model, we first use traditional AP (average precision) to evaluate the performance, and for a more intuitive understanding of the relationship with the global inference model, we also use cosine similarity to evaluate it by assembling it into the vertical distribution.

Experiment results on the spatial distribution inference are shown in Table 3. Equipped with consistency loss, both local and global branches get better performance. Besides, fusing two distributions can get better performance than a single distribution (Fusion > Global,Local). Performance of global inference model is slightly better that local one, implying higher noise level of the local inference.

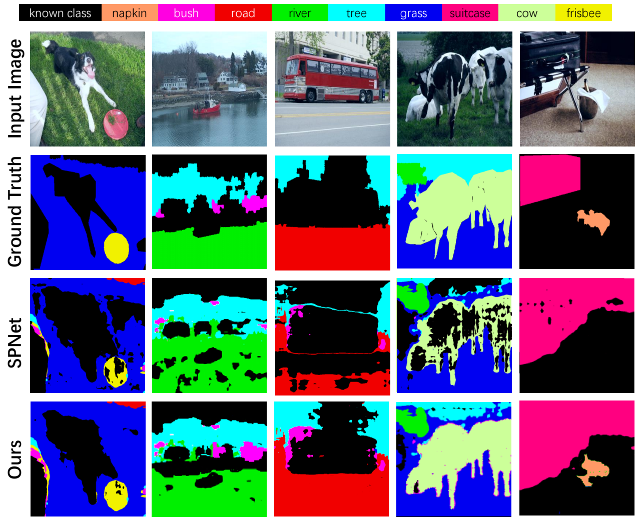

Figure 6: Qualitative results of baseline SPNet and our method on COCO-stuff.

Method	ZSL	GZSL		
		S	U	H
SPNet[49]	35.9	30.4	5.2	8.8
Context Entailment	37.9	30.8	5.7	9.6
CE with Spatial Information	38.7	30.9	6.3	10.5
CE with fused Spatial Information	**39.1**	**31.5**	**6.6**	**10.9**

Table 4: Segmentation results with estimated spatial distribution on zero shot learning (ZSL) and generalized zero shot learning (GZSL) setting. For GZSL, we report the mIoU on seen classes (S) and unseen classes (U) and the harmonic mean (H) of them.

We also evaluate the impact of spatial information on the performance of segmentation model in both zero shot and generalized zero shot setting. The experimental results on COCO-stuff are shown in Table 4. With the spatial distribution information, the performance of segmentation model in both ZSL and GZSL can be improved. And the fused spatial distribution information can get even higher performance. Some qualitative results are presented in Figure 6. Trained from the annotation provided by our Cap2Seg inference engine, the segmentation model can give a better performance on objects and stuff classes, such as cow, road, grass etc. Besides, it can also segment some hard cases on which the baseline model fails to recognize the target object as shown in the rightmost column about the napkin class.

5.3 Evaluations for Different Splits

For fairness, we also evaluate the proposed method on the class split proposed in [49]. The experimental results are found in Table 5. As stated in priro sections, the unknown classes of the class split of [49] contains some classes frequently explicitly mentioned by the image captions. This arguably explains the slight performance improvement compared with the results obtained with the split we constructed.

Method	ZSL	GZSL		
		S	U	H
SPNet[49]	35.2	27.0	9.0	13.5
Our Method	**41.5**	**27.1**	**11.3**	**16.0**

Table 5: Comparison of image segmentation performances on the class split proposed by SPNet [49] in terms of mIoU on COCO-stuff. Both settings of ZSL and general ZSL are studied.

	ZSL	plant	sheep	sofa	train	tv	U	S	H
ZSVM[21]	35.6	-	-	-	-	-	-	-	-
SPNet[49]	49.6	3.2	20.9	12.4	38.3	1.7	15.3	76.4	25.5
Our Method	**51.4**	**7.4**	**30.3**	**15.2**	**51.8**	**2.2**	**21.3**	**76.5**	**33.3**

Table 6: Experimental results on the Pascal VOC (mIoU).

Method	S	U	H
SPNet[49]	50.0	18.6	27.1
ZS3[6]	47.3	24.2	32.0
Our Method	**50.4**	**25.6**	**34.0**

Table 7: On PASCAL VOC, comparisons on the class split where unknown classes are: *cow, motorbike, airplane, sofa, cat, tv.*(mIoU).

5.4 Experimental Results On Pascal VOC

We also conduct experiments on PASCAL VOC 12. Since only 906 training images are equipped with caption annotations, we firstly train the segmentation model in the rest of training images with only the known class annotations. Then we fine-tune the segmentation model on the small part of images with unknown class information inferred by our inference engine. Provided the limited number of captions, we adopot the inference engine pre-trained on COCO dataset to infer on PASCAL. It is notable that our inference engine does not require any pre-defined class. One can construct a new CEQ with the textual representation of any new unknown class. Thus, it can be trivially transferred between different datasets.

Table 6 shows the results on the class split proposed by [49] and [21]. Our method surpass other baselines in both ZSL and GZSL. We also conduct experiments on the class split proposed by [6], shown in Table 7. Our model also surpass other methods. The experiment results show the effectiveness of our proposed Cap2Seg even with only a small number of caption annotation.

6 CONCLUSIONS

We propose a novel three branch inference model which makes full use of provided annotations, including image captions and known class annotations, to infer the occurrence probability and spatial distribution of unknown classes. These information is further aggregated with a spatial-weighted pair-wise ranking loss to give supervision on the segmentation task. In the future, we will further verify the benefit of our proposed method on different tasks, like object detection. We will also integrate the proposed inference and downstream models into a more efficient end-to-end framework.

Acknowledgement: This work is supported by BOE Technology Group Co., Ltd., Beijing Natural Science Foundation (Z190001), and National Natural Science Foundation of China (61772037).

REFERENCES

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. 2016. Label-Embedding for Image Classification. *TPAMI* 38, 7 (2016), 1425–1438.

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

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *TPAMI* 39, 12 (2017), 2481–2495.

[4] Amy L. Bearman, Olga Russakovsky, Vittorio Ferrari, and Fei-Fei Li. 2016. What's the Point: Semantic Segmentation with Point Supervision. In *ECCV*. 549–565.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*. 41–48.

[6] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. 2019. Zero-Shot Semantic Segmentation. In *NIPS*.

[7] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. 2018. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*. 1209–1218.

[8] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2016. Synthesized Classifiers for Zero-Shot Learning. In *CVPR*. 5327–5336.

[9] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In *ECCV*. 52–68.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *TPAMI* 40, 4 (2018), 834–848.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *TPAMI* 40, 4 (2018), 834–848.

[12] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR* abs/1706.05587 (2017).

[13] Ido Dagan and Oren Glickman. 2004. PROBABILISTIC TEXTUAL ENTAILMENT: GENERIC APPLIED MODELING OF LANGUAGE VARIABILITY.

[14] Jifeng Dai, Kaiming He, and Jian Sun. 2015. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *ICCV*. 1635–1643.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 4171–4186.

[16] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (2015), 98–136.

[17] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*. 2121–2129.

[18] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. 2018. Recent Advances in Zero-Shot Recognition: Toward Data-Efficient Understanding of Visual Content. *IEEE Signal Process. Mag.* 35, 1 (2018), 112–125.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[20] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *CoRR* abs/1612.03651 (2016).

[21] Naoki Kato, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2019. Zero-Shot Semantic Segmentation via Variational Mapping. In *ICCV Workshops*.

[22] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[23] Elyor Kodirov, Tao Xiang, and Shaogang Gong. 2017. Semantic Autoencoder for Zero-Shot Learning. In *CVPR*. 4447–4456.

[24] Alexander Kolesnikov and Christoph H. Lampert. 2016. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *ECCV*. 695–711.

[25] Suha Kwak, Seunghoon Hong, and Bohyung Han. 2017. Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network. In *AAAI*. 4111–4117.

[26] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *TPAMI* 36, 3 (2014), 453–465.

[27] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. 2017. Zero-Shot Recognition Using Dual Visual-Semantic Mapping Paths. In *CVPR*. 5207–5215.

[28] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. 2016. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *CVPR*. 3159–3167.

[29] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2017. Feature Pyramid Networks for Object Detection. In *CVPR*. 936–944.

[30] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. 2018. Generalized Zero-Shot Learning with Deep Calibration Network. In *NIPS*. 2009–2019.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. 3431–3440.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 3111–3119.

[33] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*.

[34] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille. 2015. Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation. *CoRR* abs/1502.02734 (2015).

[35] Pedro H. O. Pinheiro and Ronan Collobert. 2015. From image-level to pixel-level labeling with Convolutional Networks. In *CVPR*. 1713–1721.

[36] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*. 139–147.

[37] Bernardino Romera-Paredes and Philip H. S. Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*. 2152–2161.

[38] Anirban Roy and Sinisa Todorovic. 2017. Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image Segmentation. In *CVPR*. 7282–7291.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. [n.d.]. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* ([n. d.]).

[40] Johann Sawatzky, Debayan Banerjee, and Juergen Gall. 2019. Harvesting Information from Captions for Weakly Supervised Semantic Segmentation. *CoRR* abs/1905.06784 (2019).

[41] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-Shot Learning Through Cross-Modal Transfer. In *NIPS*. 935–943.

[42] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized Zero-Shot Learning via Synthesized Examples. In *CVPR*. 4281–4289.

[43] Vinay Kumar Verma and Piyush Rai. 2017. A Simple Exponential Family Framework for Zero-Shot Learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part II*. 792–808.

[44] Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 2019. *NeurIPS*.

[45] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM TIST* 10, 2 (2019), 13:1–13:37.

[46] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. 2018. Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation. In *CVPR*. 7268–7277.

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[48] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele. 2016. Latent Embeddings for Zero-Shot Classification. In *CVPR*. 69–77.

[49] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. 2019. Semantic Projection Network for Zero- and Few-Label Semantic Segmentation. In *CVPR*. 8256–8265.

[50] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. *TPAMI* 41, 9 (2019), 2251–2265.

[51] Jia Xu, Alexander G. Schwing, and Raquel Urtasun. 2015. Learning to segment under various forms of weak supervision. In *CVPR*. 3781–3790.

[52] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. 2019. Cap2Det: Learning to Amplify Weak Caption Supervision for Object Detection. In *ICCV*.

[53] Meng Ye and Yuhong Guo. 2017. Zero-Shot Classification with Discriminative Semantic Representation Learning. In *CVPR*. 5103–5111.

[54] Li Zhang, Tao Xiang, and Shaogang Gong. 2017. Learning a Deep Embedding Model for Zero-Shot Learning. In *CVPR*. 3010–3019.

[55] Ziming Zhang and Venkatesh Saligrama. 2015. Zero-Shot Learning via Semantic Similarity Embedding. In *ICCV*. 4166–4174.

[56] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*. 2921–2929.

[57] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. 2018. Weakly Supervised Instance Segmentation Using Class Peak Response. In *CVPR*. 3791–3800.