Concept Mask: Large-Scale Segmentation from Semantic Concepts

Yufei Wang[1*][0000−0002−0729−908X], Zhe Lin[2][0000−0003−1154−9907], Xiaohui Shen[3*], Jianming Zhang[2][0000−0002−9954−6294], and Scott Cohen[2]

[1] Facebook Research, Menlo Park, CA, USA
yufei22@fb.com
[2] Adobe Research, San Jose, CA, USA
{zlin,jianmzha,scohen}@adobe.com
[3] ByteDance AI Lab, Menlo Park, CA, USA
shenxiaohui@bytedance.com

Abstract. Existing works on semantic segmentation typically consider a small number of labels, ranging from tens to a few hundreds. With a large number of labels, training and evaluation of such task become extremely challenging due to correlation between labels and lack of datasets with complete annotations. We formulate semantic segmentation as a problem of image segmentation given a semantic concept, and propose a novel system which can potentially handle an unlimited number of concepts, including objects, parts, stuff, and attributes. We achieve this using a weakly and semi-supervised framework leveraging multiple datasets with different levels of supervision. We first train a deep neural network on a 6M stock image dataset with only image-level labels to learn visual-semantic embedding on 18K concepts. Then, we refine and extend the embedding network to predict an attention map, using a curated dataset with bounding box annotations on 750 concepts. Finally, we train an attention-driven class agnostic segmentation network using an 80-category fully annotated dataset. We perform extensive experiments to validate that the proposed system performs competitively to the state of the art on fully supervised concepts, and is capable of producing accurate segmentations for weakly learned and unseen concepts.

Keywords: semantic segmentation, large-scale segmentation, semi-supervised learning, weakly-supervised learning, zero-shot learning

1 Introduction

Image segmentation has attracted a lot of attention in the recent years, and has achieved great progress with the success of Deep Neural Networks (DNN) [31, 25, 1, 2]. Two popular tasks of segmentation problems are semantic segmentation and instance segmentation. Existing semantic segmentation or scene parsing methods mostly consider a small number of classes and their extension to a

* The work was done when the authors were in Adobe Research.

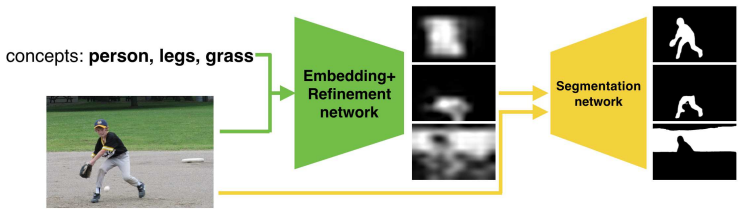

Fig. 1. Overall architecture of the proposed framework. Given a concept (can be object, object parts, stuff, etc.) and an input image, our embedding network and the subsequent attention refinement network predict a low resolution attention map, and a label agnostic segmentation network takes the attention map and the original image as input to predict a segmentation mask for the concept.

large number of classes is challenging. The main difficulty comes from arising overlap between labels when the number of labels significantly increases: for the large-scale setting, labels at different levels or different branches in the WordNet hierarchy could have complex spatial correlations and subsequently confuse the pixel level annotation tasks. For example, for the face of a person, both the fine level annotation of "face" and the higher level annotation of "person" are correct, and for the area of "clothing" on a human body can also be annotated as "person" or "body". This will cause a substantial challenge in training and evaluation of segmentation algorithms. On the other hand, pixel wise annotation for a large number of images and labels takes a lot of manual effort and is costly to obtain, and current publicly available benchmark datasets only have a small number of classes (for example, the MIT Scene Parsing Benchmark, the largest scene parsing dataset, contains annotations for only 150 classes).

As for instance segmentation, the problem only focuses on objects, and the state-of-the-art bounding box proposal-based methods [22, 8, 11] cannot handle object parts, stuff or other concepts like visual attributes. This is because the region proposal network predicts objectness of a bounding box, and naturally takes object parts or stuff as negative examples.

In this work, we take a step forward and propose a new approach for large-scale semantic segmentation. To overcome the label ambiguity issue, we formulate the task as a problem of image segmentation given an arbitrary semantic concept. For example, the concept can refer to an object, object part, object group, stuff, attribute, etc. By this formulation, we alleviate the issue of label confusion in large-scale semantic segmentation and scene parsing, which makes training and evaluation of segmentation algorithms more well-defined.

However, there is no available dataset for large-scale segmentation. To leverage the existing datasets with different levels of supervision, we use four datasets for training: a 6M Stock dataset (crawled from a stock website) with 18K image level labels; a curated a 750-concept dataset from Open Images [18] and Visual Genome [19], with bounding box annotation; MS-COCO [24] with full segmentation annotation for 80 object classes. In order to evaluate the model's capability

on weakly supervised learning, we select a diverse set of 50 test concepts among 18K concepts excluding those 750 concepts with the bounding box annotations.

Given the datasets, we propose a new weakly and semi-supervised learning approach which can leverage all the available training data in an incremental learning framework. The proposed incremental learning framework consists of three steps. First, we train a deep neural network on the stock dataset[4] to learn large-scale visual-semantic embedding between images and 18K concepts [34]. By running the embedding network in a fully convolutional manner, we can compute a coarse attention (heat) map for any given concept. Next, we attach two fully connected layers to the embedding network and fine-tune the refinement network in low resolution using the 750-concept dataset with bounding box annotations to obtain improved attention maps. We use multi-task training to learn from the new 750-concept supervision without affecting the previously learned knowledge on 18K concepts. Finally, we train a label-agnostic segmentation network which takes the attention map and original image as input and predicts a high-resolution segmentation mask without much knowledge of the concept of interest. The segmentation network is trained with only 80 object categories with pixel-level supervision but we show that it generalizes well to any semantic concept, including objects, object parts, and even background stuff, due to the use of attention maps for class-agnostic segmentation.[5] During testing, we can attach the segmentation network to the attention network to form a unified feed-forward network model.

We perform extensive experiments to validate that the proposed approach performs competitively to the state of the art on fully supervised concepts, and is capable of producing accurate segmentations for weakly learned and unseen concepts. The main contributions of this paper are as follows: 1) We address the problem of large-scale semantic segmentation with a new formulation: large-scale segmentation given a concept; 2) To study this task, we construct multiple datasets with different levels of supervision, and establish performance evaluation methods; 3) We propose a new, incremental learning approach that can predict segmentation masks for a very large number of concepts, including object, object parts, and stuff; 4) We propose a novel auxiliary loss called spatial discrimination loss for discriminative segmentation training for a large number of concepts with complex semantic relationships.

2 Related Work

Fully Supervised Semantic Segmentation Semantic segmentation has made remarkable progress with the recent advancement in deep convolutional neural networks (CNN). Many CNN-based segmentation networks [36, 31, 8, 11, 22, 6]

[4] https://stock.adobe.com

[5] Note that traditional bounding box proposal-based methods with class-agnostic segmentation could easily fail to detect proposals on object parts or stuff and the bounding box-based class-agnostic segmentation module trained with object categories cannot deal with stuff categories.

perform well on datasets with a small number of labels, such as PASCAL VOC [9] with 20 object classes, ADE20K[38] with 150 stuff/object classes.

For instance aware segmentation which requires segmentation of individual object instances, methods based on region proposals [29] perform well on the COCO dataset with 80 object classes [22, 8, 11]. However, the region proposal based methods can only handle object classes, and their generation to other concepts such as object parts or stuff is not straightforward.

These methods are all fully supervised, and assume disjoint classes, which enables training segmentation networks with a discriminative soft-max loss.

Weakly/Semi-supervised Semantic Segmentation In order to reduce annotation efforts needed in fully supervised methods, weakly supervised segmentation methods have been proposed [20, 30, 3, 14, 28, 17, 7]. Image-level annotations require minimum manual effort, but methods with such annotations have a large performance gap compared to fully supervised methods; additional label types such as bounding box annotations are exploited to improve the performance. On the other hand, some works exploit complementary data from the web [14]. Those weakly supervised methods still focus on a small set of disjoint labels.

Different from those works, this paper aims to scale semantic segmentation to a very large number of categories. We make use of all the available annotations in several datasets, thus combining different levels of annotation.

One work related to our model is by Hong *et al.* [13]. Segmentation is decoupled into two tasks with two separate networks: classification and segmentation. The classification network uses image level annotation, and the segmentation network uses pixel level annotation. However, their work still focuse on a very small number of labels, and their model cannot generalize to unseen concepts.

Another recent work related to ours is by Hu *et al.* [15]. It aims at instance segmentation on a large number of categories with a small fraction of mask annotations and a large fraction of box annotations. In contrast, our work aims to segment not only objects, but also other concepts such as stuff, parts, and visual attributes (like color); our model is learned to segment concepts trained with only image-level supervisions and can even handle unseen concepts.

Zeroshot Learning For the problem of zeroshot learning, models are tested on unseen categories by transferring knowledge from the trained categories. Semantic embedding of vectors associated with class labels are obtained from object attribute labels [16, 27, 21] or word embeddings learned from linguistic tasks [10, 32, 26]. Zeroshot learning can also be applied to segmentation tasks. With the embedding network that maps an image to a word embedding space, segmentation models have the potential to generate masks given an unseen concept [35].

Large Scale Segmentation/Parsing Zhao *et al.* aim to recognize and segment objects with open vocabulary [35], which is in line with our goal of large scale segmentation. Words and images are embedded into a joint space to allow

zero-shot learning. Our work is different from theirs in that (1) Zhao *et al.* address only zero-shot segmentation while we aim to solve weakly supervised and zero-shot segmentation in a unified framework, (2) Zhao *et al.* use WordNet for modeling hierarchical label relationships while we consider more complex label relationships including spatial overlap/exclusion which makes trained segmentation models more discriminative, (3) Zhao *et al.* view the open-vocabulary scene parsing as a concept retrieval problem, whereas we assume a target concept is given as an extra input to predict the segmentation mask. Our task is easier to evaluate and with less ambiguity in the ground truth masks.

3 Dataset

Public segmentation datasets typically contain pixel-level annotations on only a small number of labels. On the other hand, datasets with a much larger vocabulary are only weakly annotated, either with bounding box or image-level labels. To make the most of the available datasets, we form a combined dataset, containing different levels of annotations:

- **COCO-80**: MS-COCO dataset [24] pixel level annotation on 80 categories.
- **OIVG-750**: Combined Open Images [18] and Visual Genome [19] dataset with 750 concepts (including COCO concepts) with bounding boxes.[6]
- **Stock-18K**: 6M Stock dataset annotated with 18K tags.

With the combined dataset, we can evaluate the performance of segmentation methods under different levels of supervision by constructing the following test set: (1) strongly supervised concepts: COCO-80 test set, (2) box-level weakly supervised concepts: Weak-Box-670, obtained by excluding 80 COCO categories from OIVG-750, (3) image-level weakly supervised concepts: Weak-Image-50, obtained by choosing 50 new concepts from OIVG excluding OIVG-750.

For testing on weakly-supervised settings, there is no available segmentation ground truth for classes outside COCO-80. Therefore we generate pseudo-ground truth from bounding boxes, using an automatic segmentation model [33]. It is then manually cleaned up, and in the supplementary material we show some examples of the generated pseudo-ground truth masks. Note that we use the pseudo-ground truth masks only for evaluation.

4 Proposed Approach

The overall framework of our large-scale segmentation system is illustrated in Figure 1. It is composed of an embedding and refinement network that produces an attention map from the input image and a specified concept, and an attention-driven label-agnostic segmentation network that predicts a final segmentation

[6] Visual Genome has more than 10000 classes, and Open Images has 545 trainable classes. We merge the labels of the two datasets, and filtered out classes with very few examples. 750 concepts containing objects, object parts, and stuff are selected.

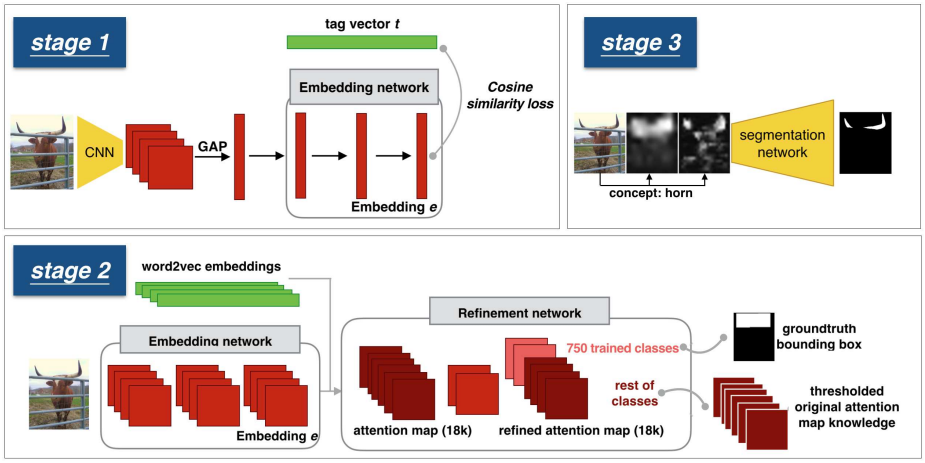

Fig. 2. Three stages of the training framework. Stage1: embedding network trained on image level annotation. Stage2: multi-task training of attention network. It finetunes the embedding network while training the refinement network from scratch. It refines the attention map on 750 concepts with bounding box supervision, meanwhile preserves the knowledge learned from embedding network on 18K concepts. Stage3: label agnostic segmentation network that takes the original image and two refined attention maps (generated from input image with two scales), and predicts the segmentation mask.

mask. Utilizing three different levels of supervision, we train the entire framework incrementally in three stages:

1. Train an embedding network on Stock-18K that learns the visual-semantic embedding between images and 18K semantic concepts. Only image-level annotations are used in this stage. After training, we transform the network to fully convolutional, which can generate a low resolution attention map given an input image and any of the 18K concepts.
2. Append a refinement module to the end of the embedding network, and train the refinement network on OIVG-750 with bounding box annotations, aiming at obtaining attention maps of higher quality.
3. Train a label agnostic segmentation network on COCO-80 with 80-class full segmentation supervision. The network takes the initial attention maps together with the image as input, and predicts a higher resolution segmentation mask with more accurate boundaries for the concept.

4.1 Embedding Network

We first utilize the Stock-18K dataset with image-level annotations to learn large-scale visual-semantic embedding. The dataset has 6 million images, each with heavily annotated tags from an 18K vocabulary. The training set is denoted as $\mathcal{D} = \{(I, (w_1, w_2, ..., w_n)\}$, where I is an image and w_i is the word vector representation of its associated ground-truth tags.

Word Embedding Instead of using off-the-shelf word embeddings trained on a text corpus, we use point-wise mutual information (PMI) to learn our own word embeddings for each tag w in the vocabulary. PMI is a measure of association commonly used in information theory and statistics [5]. We follow [4] to calculate PMI matrix and then do eigenvector decomposition to the matrix to get the word vector. More details are shown in the supplementary material.

Since each image is associated with multiple tags, in order to obtain a single word vector representation of each, we calculate a weighted average over all the associated tags: $t = \sum_{i=1}^{n} \alpha_i w_i$ where $\alpha_i = -\log(p(w_i))$ is the inverse document frequency (idf) of the word w_i. We call the weighted average *soft topic embedding*.

Joint Word-Image Embedding The embedding network is learned to map the image representation and the word vector representation of its associated tags into a common embedding space. As shown in Figure 2 stage 1, each image I is passed through a CNN feature extractor. Here we use ResNet-50 [12] as feature extraction network. After global average pooling (GAP), the visual feature is then fed into a 3-layer fully connected network, denoted as **Embedding network**, with each fc-layer followed by a batch normalization layer and a ReLU layer. The output is the visual embedding $e = embed\_net(I)$, and is align with the soft topic word vector t by a cosine similarity loss: $L_{embed}(e, t) = 1 - \frac{e^T t}{\|e\|\|t\|}$.

Attention Map After the embedding network is trained, to predict an attention map for a given concept, we remove the global average pooling layer, and transform the network to a fully-convolutional network by converting the fully connected weights to 1×1 convolution kernels and the batch normalization layers to spatial batch normalization layers. After this transformation, we can obtain a dense embedding map given an image and a word vector, in which the value at each location is the similarity between the word and the image region around that location. Thus the embedding map can also be viewed as an attention map for that word. Note that the way we generate the attention map is similar to [37]. However, we use soft topic embedding instead of discriminative classification so the attention map has better spatial coverage than the one in [37].

Formally, the attention map for a given concept w can be calculated as:

$$\alpha^0_{(i,j)} =< e_{i,j}, w > \tag{1}$$

where (i, j) is the location index for the attention map. For an unseen concept that is not used in our image-word embedding training, as long as we can obtain its word vector w, we can still obtain its attention map using Eqn.1. Therefore, our embedding network can be generalized to any arbitrary concept.

4.2 Attention Map Refinement

Although the embedding network trained on image level annotation can predict attention maps for any given word vector, the quality of the attention maps is still very coarse due to the lack of annotations with spatial information.

In order to improve the quality of the attention map, we leverage existing finer-level annotations, namely the object bounding box annotations that are available in several large-scale datasets. Specifically, we use the OIVG-750 dataset to train a network for attention map refinement.

Refinement Network Architecture As shown in Figure 2 stage 2, the refinement network is appended at the end of the embedding network, and is composed of two convolutional layers with 1×1 kernels followed by a sigmoid layer. By treating word embeddings as convolutional kernels, embedding network can now output 18K coarse attention maps. The two-layer refinement network takes those coarse attention maps as input, and learns a non-linear combination of the concepts to generate refined attention maps for the 750 classes. This encourages the refinement network to consider relationships between concepts during training.

Multi-task Training For a given concept, training signal for its attention map is a binary mask based on the ground-truth bounding boxes, and a sigmoid cross entropy loss is used. Embedding network is also finetuned for better performance. However, since the bounding box annotations are only available for the 750 concepts, if we only train the network on those classes, the previously learned attention maps for the rest of 18K concepts will be corrupted if we also finetune the embedding network layers. Inspired by [23] on learning without forgetting, in order to preserve the learned knowledge from the rest of 18K concepts, an additional matching loss is added: the original attention maps generated by the embedding network are binarized with a threshold, and sigmoid cross entropy loss is exerted for the refined attention maps to match the original attention maps. The multi-task loss function is therefore as follows:

$$L = L_{xe}(G, \alpha) + c \sum_{k \in \Psi_N} L_{xe}(\mathrm{B}(\alpha_k^0), \alpha_k) \tag{2}$$

where $L_{xe}(p, q)$ is the cross entropy loss between true distribution p and predicted distribution q. α is the attention map of the given concept, G is the ground truth mask with 1 being inside the bounding box, 0 outside. $\mathrm{B}(\alpha)$ is the binary mask after thresholding the attention map. α_k^0 and α_k are original attention map and refined attention map respectively. Ψ_N is the set of indices of top N attention maps with the highest activation. The matching loss is exerted on attention maps with high activation only to avoid bias toward irrelevant concepts. c is the weight balancing the losses. We choose $N = 800$, and $c = 10^{-6}$.

Spatial Discrimination Loss The reason we used sigmoid cross entropy loss during training instead of softmax loss as in semantic segmentation is that there are many concepts whose masks are overlapping with each other. It is especially common for objects and their parts. For example, the mask of face is always covered by the mask of person. Using softmax loss therefore would discourage the mask predictions on those concepts one way or another. At the same time, there

are still many cases where the masks of two concepts never overlap. To utilize such spatial relationships between label pairs and make training of the attention maps more discriminative, we propose a novel auxiliary loss for discriminating those spatially non-overlapping concepts, referred to as spatial discriminative loss, to discourage high responses for spatially conflicting concepts occuring at the same time.

In particular, we calculate the mask overlap ratio between every co-occurred concept pair in the training data:

$$O(i,j) = \frac{\sum_n |a_n(i) \cap a_n(j)|}{\sum_n |a_n(i)|} \tag{3}$$

where $a_n(i)$ is the mask of the i-th concept in image n, and $|a_n(i) \cap a_n(j)|$ is the overlapping area of between concepts i and j. Here image n has to include both i and j concepts to avoid impact of incomplete annotation. Note that the mask overlap ratio is non-symmetric. In the supplementary material, we show a subset of the overlap ratio matrix $O(i,j)$.

With the overlap ratio matrix $O(i,j)$, a training example of a concept i can serve as a negative training example of its non-overlapping concept j, i.e., for a particular location in the image, the output for concept j should be 0 if the ground-truth for concept i is 1. To soften the constraint, we further weight the auxiliary loss based on the overlap ratio, where the weight γ is calculated as:

$$\gamma_{ij} = \begin{cases} 1 - O(i,j), \text{ if } & O(i,j) < 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

4.3 Label Agnostic Segmentation Network

Our attention map refinement network now can predict low resolution attention map for an arbitrary concept using its word vector representation. To further obtain the mask of the concept with higher resolution and better boundary quality, we train a label agnostic segmentation network that takes the original image and the attention map as input, and generates a segmentation mask without knowing the concept, as shown in Figure 2 stage 3. Since the goal of the segmentation network is to generate foreground segmentation mask given the prior knowledge of attention map, the segmentation network can generalize to unseen concepts, even though it is entirely trained on COCO-80 with only 80 object classes.

To segment the masks at different scales, we generate multiple attention maps by feeding the embedding network with different input image sizes (300 and 700 in our experiments). The resultant attention maps are then upsampled to serve as the extra input channel to the segmentation network along with the image.

To make the segmentation network focus on generating accurate masks instead of having the extra burden of predicting the existence of the concept in the image, we normalize the attention maps to $[0, 1]$. We found that such training strategy can learn better segmentation networks. During testing, the attention maps are normalized in the same way, and the verification of the existence of the concept is done separately, with details presented in the supplementary material.

For the architecture of segmentation network, we use an architecture that extracts and combines high level and low level features to predict a concept mask with accurate boundaries. See the supplementary materials for more details.

4.4 Weakly Supervised Segmentation

During testing stage, for the 18K concepts that are only trained with image level supervision, we do not directly use the attention map from the refinement network for that concept as the input to the segmentation network. This is because the segmentation network only sees the examples of the COCO-80 during training, which has the attention map trained with bounding box / pixel-wise segmentation supervision. Thus, the discrepancy between the lower-quality attention maps of the 18K concepts and the higher-quality attention map of the 750 concepts will impact the segmentation performance on 18K concepts.

Therefore, for a concept q from the 18K concepts with image level supervision, we find its nearest neighbor concept p in the embedding space from the 750 concepts, and the attention maps is a linear combination $\alpha = \theta\alpha_q + (1-\theta)\alpha_p$ of the attention maps from the two concepts, with θ decided on validation set.

5 Experiments

In this section, we provide visual and numerical results on attention map prediction and segmentation mask generation under different levels of supervision. Experimental details are shown in the supplementary material.

5.1 Datasets

For COCO-80, we use the train2014 split, with 80k training images. For OIVG-750, there are 540k training images, and training examples for each concept varies from 8 to 100k. Stock-18K has 6M training images, and 30 tags for each image on average. Test/validation set for OIVG-750, Weak-Box-670, Weak-Image-50 have 5-10 examples per concept, and one example is held for validation.

In Figure 3, we show the example images and ground truth labels from different datasets. Each row shows an example of a dataset with annotation. In the last column, we also show an example test image from Weak-Image-50, for which the annotation is the pseudo-ground-truth mask.

5.2 Attention Map Evaluation

For attention map generation, we use several ways for evaluation. Following [34], we use Pointing Game for evaluation. For an image, if the maximum point in attention map lies in the ground truth mask, a hit is counted. We can measure the mean accuracy across all the concepts. We can also use IOU for evaluation.

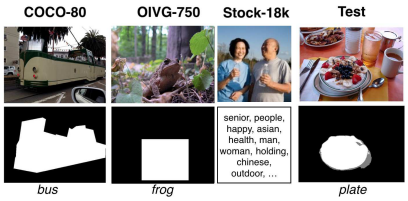

Fig. 3. Examples of our datasets with different levels of annotation. First row is the original image, and second row shows the annotation labels. The Stock-18k image is from **arekmalang - stock.adobe.com**.

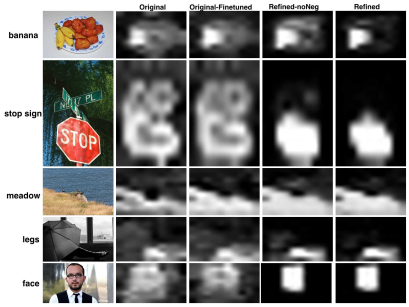

Fig. 4. Examples of attention map generated from different models/phases.

The attention map is a probability mask that ranges from 0 to 1, and we calculate IOU as follows:

$$\text{IoU}_n = \frac{\alpha_n * G_n}{\max(\alpha_n, G_n)} \tag{5}$$

where n is image index, α is the attention map, and G is the ground truth. When there is only bounding box ground truth available, we use the pseudo-ground-truth for evaluation.

Table 1. Performance of different models on attention map generation

	Original	Original-Finetuned	Refined-noNeg	Refined
Pointing Game	0.578	0.631	0.806	**0.810**
IoU	0.262	0.288	0.416	**0.421**

In Table 1, we compare the performance of different models/phases on attention map generation, using OIVG-750 evaluation set. **Original** is the original attention map we obtain from the embedding network, as described in Section 4.1. **Original-Finetuned** is the attention map from the embedding network, after finetuning with the refinement network. **Refined-noNeg** is the result from the refinement network as described in Section 4.2, without using the negative examples from non-overlapping concept. **Refined** is our full model.

In Figure 4, we also show the visual result of attention map generated from different models. We can see that the original attention map already generates acceptable attention maps, but it is noisy, and sometimes locating to objects when the concept is object part (see example of *face*). The refined attention map is much cleaner, and is covering the whole object/stuff. The comparison between the result from Refined-noNeg and Refined shows that by using negative samples from non-overlapping concept, the attention map is cleaner visually, and is more discriminative.

5.3 Segmentation Evaluation

In this section, we evaluate our model with different levels of supervision, and compare our results with different baselines. For quantitative evaluation, we simply binarize the soft segmentation outputs using the threshold of 0.5 for all models. Given the binary ground truth mask and prediction mask for one concept, we can calculate precision/recall/IoU. Over different concepts, we calculate mean precision, recall, and mean IoU.

Dataset Our model uses different levels of supervision, and thus we can evaluate the model performance on concepts with different levels of supervision separately: categories inside **COCO-80** are used to train the attention network and segmentation network, therefore we can evaluate it for full (strong) supervision. We use 5000 miniVal2014 split for evaluation, and all the annotated concepts in the images are evaluated; **Weak-Box-670** is used to evaluate segmentations with bounding box-level supervision; **Weak-Image-50** is used to evaluate our model's performance on concepts with only image-level supervision.

Results We compare the performance of our model and baselines on different levels of supervision in Table 2 and Table 3.

Table 2. Comparison of the performance of our model and baselines, on different levels of supervision. On the left, we show the descriptions of different baselines we use

Model	Description
FCIS	Semantic segmentation trained and tested on COCO[34]. The model generates instance mask for a given concept, and we merge all the instances of one concept to one binary mask.
Mask R-CNN	State-of-the-art semantic segmentation model[11].
Saliency-DSS	The state-of-the-art saliency network for a given image [15].
Mask R-CNN*	Our modified version of Mask-RCNN to better perform on our task. See main paper for details.

Test Dataset	Model	Precision	Recall	IoU
COCO-80	FCIS[34]	**0.864**	0.596	0.442
	Mask R-CNN[11]	0.854	0.668	**0.525**
	Saliency-DSS[15]	0.286	0.492	0.181
	Ours	0.656	**0.722**	0.402
Weak-Box-670	Saliency-DSS[15]	0.592	0.407	0.307
	Mask R-CNN*	0.281	0.111	0.089
	Ours	**0.741**	**0.800**	**0.609**
Weak-Image-50	Saliency-DSS[15]	0.387	0.493	0.283
	Mask R-CNN*	0.268	0.147	0.133
	Ours	**0.580**	**0.539**	**0.412**

Table 3. Ablation study for our model on different levels of supervision. On the left we show the descriptions of the models we compare with

Model	Description
Ours-noNeg	Our model without negative training examples from non-overlapping concepts as described in Section 4.2.
Ours-singleAtt	The segmentation network takes only one attention map from image size 300×300 as input.

IoU	Ours	Ours-noNeg	Ours-singleAtt
COCO-80	**0.402**	0.395	0.363
Weak-Box-670	**0.609**	0.594	0.599
Weak-Image-50	**0.412**	0.356	0.390

As shown in Table 2, the performance of our model decreases with less supervision. For the fully supervised concepts, our method performs competitively with semantic segmentation model FCIS, with 4% of gap, and has moderate gap with curent state-of-the-art semantic segmentation model Mask R-CNN. The gap is predictable, because our model handles not only 80 categories inside COCO-80, but also orders of magnitude more concepts outside COCO. The closeness between the two models shows that although our model aims at a much bigger concept set, it performs very well on COCO concepts.

For the weakly supervised setting, we first compare our method with the saliency baseline. Since the concepts for evaluation are manually picked and the images used for evaluation are manually filtered for insuring the quality of the groundtruth, one concern is that the test set is not sufficient to test concept segmentation, and a saliency object detection is enough. Here by showing the performance of the saliency detection model is poor, we demonstrate that our test dataset is valid for evaluating our task.

We also train a modified Mask-RCNN (notated as Mask-RCNN*) for weakly supervised results. The original Mask R-CNN [11] predicts a mask independently for each of the 80 COCO classes. For an RoI associated with ground-truth class k, loss is defined as per pixel sigmoid loss only on the kth class. However, this does not apply to our problem, because there is no mask annotation on our large scale OIVG-750 dataset. Therefore, we modify the segmentation head to predict a label-agnostic mask, which is trained only on COCO-80. Mask-RCNN* does not perform well, and the reason can be summarized as follows: First, Mask-RCNN cannot handle stuff classes, such as sky, tree, etc., because the segmentation head only sees 80 object classes with bounding boxes. In contrast, our two-stage model does not rely on bounding boxes, so it can handle stuff very well even though the segmentation head is only trained on 80 object classes. Second, for the large number of classes (750) and highly overlapping concepts, Mask-RCNN has a very low box detection rate, which might be due to conflict of bounding box proposals among object parts, stuff, object classes.

For COCO-80, the IoU for the saliency detection result is very low, whereas for the other test dataset, the saliency performance is higher than that in COCO-80. This is due to the distributions of concepts in OIVG and COCO are essentially different, the former contains more cases with larger objects/stuff. Despite the higher performance on those test sets, we still see a great improvement of our method over saliency.

In Table 3, we show an ablation study of our approach with respect to the final performance. Our full model outperforms Ours-noNeg and Ours-singleAtt on all three cases of supervision setting, indicating the necessity of negative examples and the multi-scale attention map input to the segmentation network.

Figure 5 shows visual examples of our segmentation result. For object, object part, and stuff, we show three examples from each type. All three object categories are from COCO-80. *jean* and *frond* are from Weak-Image-50. The rest four categories are from Weak-Box-670. We also show different baselines' results. We provide more qualitative results and failure cases in the supplementary.

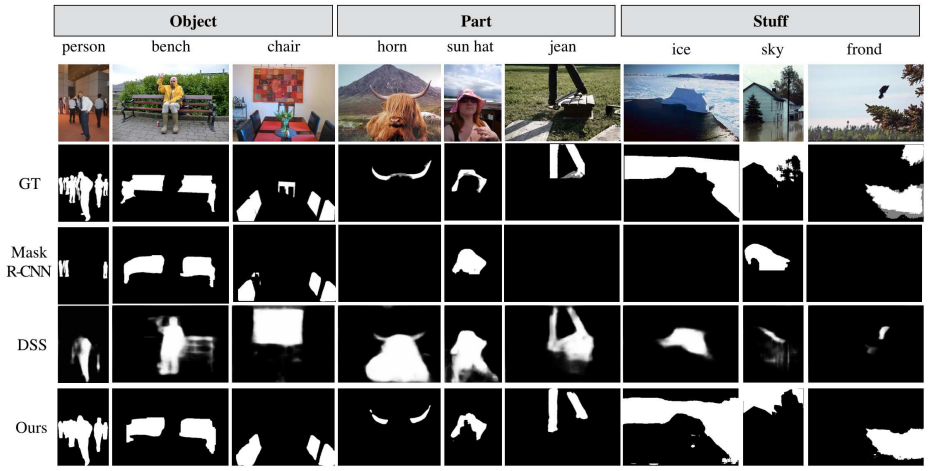

Fig. 5. Example of visual result of our segmentation network.

5.4 Zero-shot Learning

Since we use an embedding network for attention map prediction, with the word embeddings, our model can potentially handle unseen concepts. To test this potential, we curate 10 concepts outside the 18K concepts that our model is trained on, each with 5-10 test examples. The IoU of our method is 0.436, and the IoU of Saliency-DSS is 0.298. Further study with a larger test set is needed to fully justify the zero-shot learning ability of our model.

6 Conclusion

In this paper, we study semantic segmentation at a very large scale. With a large number of labels, training and evaluation of segmentation models are very challenging due to complex correlations between labels. To address the issue, we formulate the problem as conditional image segmentation given a semantic concept. Under this formulation, we propose a powerful weakly and semi-supervised segmentation framework that can handle a large number of concepts including objects, parts, stuff, attributes, and even unseen concepts. The framework consists of three parts: 1) an embedding network that maps the image and a large scale of concepts into the same space; and 2) an attention network which refines the embedding network to predict low resolution attention maps; and 3) a label agnostic segmentation network which generates segmentation masks given the attention map of a concept. Experiments show that our system performs competitively to state-of-the-art semantic segmentation models on concepts with full supervision, and is able to generate segmentation results for a large number of concepts with different levels of weak supervision, and even for unseen concepts.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
2. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
3. Chen, X., Shrivastava, A., Gupta, A.: Enriching visual knowledge bases via object discovery and segmentation. In: CVPR. pp. 2035–2042. IEEE Computer Society (2014)
4. Chollet, F.: Information-theoretical label embeddings for large-scale image classification. CoRR **abs/1607.05691** (2016)
5. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (Mar 1990), http://dl.acm.org/citation.cfm?id=89086.89095
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223. IEEE Computer Society (2016)
7. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. pp. 1635–1643 (2015)
8. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: CVPR. pp. 3150–3158. IEEE Computer Society (2016)
9. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (Jun 2010), http://dx.doi.org/10.1007/s11263-009-0275-4
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS. pp. 2121–2129 (2013)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV. pp. 2980–2988. IEEE Computer Society (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778. IEEE Computer Society (2016)
13. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) NIPS. pp. 1495–1503 (2015)
14. Hong, S., Yeo, D., Kwak, S., Lee, H., Han, B.: Weakly supervised semantic segmentation using web-crawled videos pp. 2224–2232 (2017)
15. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to Segment Every Thing. In: CVPR (2018)
16. Jayaraman, D., Grauman, K.: Zero-shot recognition with unreliable attributes. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) NIPS, pp. 3464–3472 (2014)
17. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV (4). Lecture Notes in Computer Science, vol. 9908, pp. 695–711. Springer (2016)
18. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages:

A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
20. Kttel, D., Guillaumin, M., Ferrari, V.: Segmentation propagation in imagenet. In: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV (7). Lecture Notes in Computer Science, vol. 7578, pp. 459–473. Springer (2012)
21. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 453–465 (Mar 2014). https://doi.org/10.1109/TPAMI.2013.140, http://dx.doi.org/10.1109/TPAMI.2013.140
22. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4438–4446. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.472, https://doi.org/10.1109/CVPR.2017.472
23. Li, Z., Hoiem, D.: Learning without forgetting. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV (4). Lecture Notes in Computer Science, vol. 9908, pp. 614–629. Springer (2016)
24. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (5). Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
25. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1520–1528. ICCV '15, IEEE Computer Society, Washington, DC, USA (2015). https://doi.org/10.1109/ICCV.2015.178, http://dx.doi.org/10.1109/ICCV.2015.178
26. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. In: International Conference on Learning Representations (ICLR) (2014)
27. Parikh, D., Grauman, K.: Relative attributes. In: Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V. (eds.) ICCV. pp. 503–510. IEEE Computer Society (2011)
28. Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1796–1804. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.209, https://doi.org/10.1109/ICCV.2015.209
29. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. pp. 91–99 (2015)
30. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. pp. 1939–1946 (2013)
31. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
32. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS. pp. 935–943 (2013)
33. Xu, N., Price, B.L., Cohen, S., Yang, J., Huang, T.S.: Deep grabcut for object selection. CoRR **abs/1707.00243** (2017), http://arxiv.org/abs/1707.00243

34. Zhang, J., Lin, Z.L., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV (4). Lecture Notes in Computer Science, vol. 9908, pp. 543–559. Springer (2016)
35. Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. In: ICCV. pp. 2021–2029. IEEE Computer Society (2017)
36. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 6230–6239. IEEE Computer Society (2017)
37. Zhou, B., Khosla, A., Lapedriza, ., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929. IEEE Computer Society (2016)
38. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 5122–5130. IEEE Computer Society (2017)