Weakly Supervised Segmentation with Maximum Bipartite Graph Matching

Weide Liu
Nanyang Technological University, Singapore
weide001@e.ntu.edu.sg

Chi Zhang*
Nanyang Technological University, Singapore
chi007@e.ntu.edu.sg

Guosheng Lin†
Nanyang Technological University, Singapore
gslin@ntu.edu.sg

Tzu-Yi HUNG
Delta Research Center
tzuyi.hung@deltaww.com

Chunyan Miao
Nanyang Technological University, Singapore
ascymiao@ntu.edu.sg

ABSTRACT

In the weakly supervised segmentation task with only image-level labels, a common step in many existing algorithms is to first locate the image regions corresponding to each existing class with the Class Activation Maps (CAMs), and then generate the pseudo ground truth masks based on the CAMs to train a segmentation network in the fully supervised manner. The quality of the CAMs has a crucial impact on the performance of the segmentation model. We propose to improve the CAMs from a novel graph perspective. We model paired images that contain common classes with a bipartite graph and use the maximum matching algorithm to locate corresponded areas in two images. The matching areas are then used to refine the predicted object regions in the CAMs. The experiments on Pascal VOC 2012 dataset show that our network can effectively boost the performance of the baseline model and achieves new state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Computer vision**.

KEYWORDS

weakly-supervised; segmentation; graph matching

ACM Reference Format:
Weide Liu, Chi Zhang, Guosheng Lin, Tzu-Yi HUNG, and Chunyan Miao. 2020. Weakly Supervised Segmentation with Maximum Bipartite Graph Matching. In *Proceedings of the 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/10.1145/3394171.3413652

*Contributed equally to this research.
†Corresponding author: G. Lin (e-mail: gslin@ntu.edu.sg)

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

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413652

1 INTRODUCTION

In the past few years, Deep Neural Networks have made remarkable breakthroughs in many computer vision tasks, such as image classification [9, 49, 63], object detection [15] and semantic segmentation [7, 34, 35, 37, 64, 65]. Due to its data-driven nature, large-scale labeled datasets are often necessary to enable the training of deep models. However, data labeling can be prohibitively expensive in tasks like semantic segmentation, as each pixel in the images must be manually annotated, which brings difficulty in applying segmentation models to real world problems. To alleviate this issue, researchers investigate various learning approaches with weak supervision to undertake segmentation tasks. Weak supervisions have cheap labeling costs and are often available in other computer vision tasks where numerous data is available. For example, weak supervisions with bounding boxes[8, 39], scribbles[32, 33], image labels[2, 3, 22, 27, 44, 46, 48, 53], and extreme points[4] annotations are widely studied in the literature.

In this paper, we focus on weakly supervised segmentation with image-level labels where only the existence of categories is available in the training images. Training segmentation models with image labels is a challenging problem set-up as a network must infer the whole object regions from classification losses. Many state-of-the-art methods are based upon a three-step baseline approach: first, locate the confident seed regions that are related to the classification prediction. Second, refine the seed regions and boundaries. Finally, use the mined regions as the pseudo ground truth to train a segmentation model in a fully supervised manner. Class Activation Maps (CAMs)[2, 3, 21, 27, 68] are widely adopted as an initial step to acquire seed regions in previous works. By inspecting the regions that contribute much to the classification output, CAMs can highlight the areas corresponding to each existing class. However, as is observed in many previous works[2, 3, 27, 62, 68], although the located regions are often locally correct, they tend to focus only on the most discriminative regions with strong semantic cues instead of the whole object regions. Therefore, directly using the high activation regions as the low-quality pseudo ground truth to train the segmentation model can not result in satisfactory results. Many approaches are proposed in the literature to refine the seed discriminative regions generated by the CAMs. For example, Wei *et al.*[55] proposes to iteratively erase the discriminative regions and

force the network to discover more object regions, Ahn *et al.* [2] utilize the inter-pixel relations to refine the object boundary. Our goal is also to optimize the output of the CAMs and generate better pseudo ground truth masks. We approach the weakly supervised segmentation problem from a novel perspective of graph theories. Specifically, we apply graph techniques that model the region relations between images with Maximum Bipartite Matching to better infer the object regions and we call our proposed network MBMNet.

As we want to find out more object regions to improve the quality of pseudo ground truth masks, we can better achieve this goal by comparing a pair of images that contain common categories. Unlike previous training protocols where the training images are used for learning independently, we train our network by sampling paired images. Our network first encodes representations of them with a parameter-shared Siamese encoder and then model their feature representations with a bipartite graph. We find the maximum bipartite matching (MBM) between the graph nodes to determine the relevant feature points in two images, which are then used to enhance the corresponded representations. Based on the enhanced feature representations, we can generate better CAMs with more object regions involved. The enhanced feature representations are also utilized to supervise predictions generated directly from the original representations, which serves as an auxiliary self-supervised loss in the training process.

To further improve object regions predictions, we can utilize the feature similarity and consistency under certain feature spaces to propagate high-level semantic information in the neighborhood. For example, in an image of a dog, it is likely that the head region contains strong semantic cues while the body does not, which may result in incomplete region predictions in the CAMs. However, the regions of the head and body may have similar responses in the color space, based on which, we can propagate the semantic information from the head regions to the body region. To this end, we first create an undirected graph with each pixel location as the graph node. Then a correlation score is generated for every pair of neighboring nodes under a learnable feature space. Based on the normalized correlation score, the node representations are updated by selectively aggregating information from the neighboring nodes. After the message propagation in neighboring nodes, more object regions can be discovered.

With the combination of the two graph techniques, our proposed network can generate better CAMs to train a segmentation model without additional information. We conduct comprehensive experiments on the PASCAL VOC 2012 [10] dataset to validate the effectiveness of our network. Our main contributions are summarized as follows:

- We propose MBMNet which solve the weakly supervised segmentation problem from a novel graph perspective.
- To find out more object regions in the CAMs, we propose to model pairwise relations between two images with a bipartite graph and use the maximum matching to discover common object regions. We then enhance the representations in these regions to generate better CAMs.
- We further propose to refine the representations in individual images by propagating high-level information between neighbouring regions based on node similarity under a learnable space.
- Experiments on the PASCAL VOC 2012 dataset show that our algorithm on both the validation and testing set outperforms the baseline methods and achieves new state-of-the-art performance.

2 RELATED WORK

Weak supervision in image segmentation. In order to handle the data deficiency problem in the image segmentation task, researchers have explored various types of weak supervision, such as bounding boxes [8, 39], scribbles [32, 33], and image labels[2, 3, 12, 22, 27, 42, 44, 46–48, 53, 66, 67]. These kinds of labels are less expensive to annotate and can be often acquired from existing large-scale datasets. The most well-studied weak supervision is the image-label supervision, which is also the most challenging set-up. There is also a line of literature investigating weak supervision for instance segmentation [2, 25, 59], where the network must not only locate the regions corresponding to each category, but also distinguish instances.

Image label as weak supervision. Image labels as the weak supervision for segmentation has been widely studied in the past few years. Many recent approaches [2, 3, 55] use CAMs to establish the relations between spatial locations and image label prediction. However, image classification network tends to focus only on the most discriminative regions which can not be directly used to supervise the training of segmentation models. To solve this problem, researchers have designed various ways to expand the object region maps from the most discriminative parts to the entire object regions. For example, Wei *et al.* [55] expands the seed regions generated with CAMs by erasing the discriminative regions detected by region-mining model and then re-train the model with the erased images. Yu *et al.* [62] enlarge object regions by fusing the different discriminative regions generated by convolutional layers with different dilation rate. Different convolutional layers are expected to capture different discriminative object parts. In comparison with previous methods, our method expands the discriminative region by incorporating the co-occurrent objects in paired images.

Researches also propose to expand the discriminative regions by seeking external techniques. For example, Seenet [17] incorporates saliency maps into the network training to better inference the background and foreground object regions. Wei *et al.* [55] proposes to mine the object regions in the iterative training process and Kolesnikov *et al.* [27] proposes to refine the boundary with crf [28]. Ann *et al.* [2] utilizes the inter-pixel relations to refine the boundaries of the object areas. There are also works [22, 46] utilizing external web images to improve the segmentation accuracy. Other methods adopting motion videos [16], instance saliency mask [11] also yield promising results.

Graph Matching. Graph matching has been widely used and studied in computer vision tasks, such as, object recognition [60] which explore the shape correspondences within a pair of images. Ye *et al.* [58] proposes a dynamic graph matching strategy to optimize the parameter to solve the video Re-ID task, and Zhou *et al.* [61] utilize an on-line and off-line transfer matching to person Re-ID task. Harchaoui *et al.* [13] uses kernel graph matching to

solve the image classification task. Zhang *et al.* [63] propose to use the Earth Mover's Distance to find the matching regions between two images for few-shot classification. Graph matching also has been used in image segmentation, such as Kainmueller *et al.* [24] and Martins *et al.* [38] segment the similar images with the graph matching. Zhang *et al.* [64] use the bipartite graph with attention mechanism to establish cross-image region correspondence for few-shot image segmentation.

Image co-segmentation. The image co-segmentation task is defined as segmenting out the common objects as foreground with a set of images. Many traditional works[23, 51] generate the foreground masks of the common objects by maximizing an energy function. Recently, many co-segmentation algorithms based on deep neural networks are proposed[6, 19, 31]. Though our network also aims to find the common object regions, we only use it as an intermediate step to enhance the feature representation and no explicit supervision of the common areas is provided in the training process.

3 METHOD

The pipeline of our framework for weakly supervised segmentation mainly includes three stages as shown in Figure. 1: the CAMs generation stage(our MBMNet), the boundary refinement stage and the fully supervised training stage. Our proposed method focuses on the first stage which aims to generate high-quality CAMs. In the second stage, we adopt the boundary refinement method proposed in [2] to further refine the object region predictions and generate the pseudo ground truth masks. Finally, we use the pseudo ground truth masks to train a segmentation network in a fully supervised manner for the final prediction.

In this section, we present our network design for CAMs generation in detail. Figure. 2 shows an overview of the network architecture. Compared with previous methods, the biggest difference in our design is that our network has a symmetric two-branch structure which generates the CAMs of two sampled images. We first describe the two key components in our network design and then provide details about the training protocols.

3.1 Maximum Bipartite Matching Module

As the goal of our design is to utilize the regions of co-occurrent objects in paired images to improve the CAMs, our network has a Siamese-like architecture that takes a pair of sampled images as input. We first encode two images into feature representations with a parameter-shared Convolution Neural Network. Then we reinforce the representation of the regions of the co-occurrent objects by re-weighting the features. In the end, the network generates the class probability for each location, which serves as the CAMs. To effectively locate the co-occurrent object regions in two images, we adopt graph techniques to match the elementary representations in two images. Concretely, we first model the two image representations with a bipartite graph $\mathbf{G} = \{\mathbf{U}, \mathbf{V}, \mathbf{E}\}$, where $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2 ... \mathbf{u}_{H\times W}\}$ and $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2 ... \mathbf{v}_{H\times W}\}$ are the two vertex sets which are constructed by the feature vectors of all locations from two images, and $\mathbf{E}$ is the edge set that connects the vertexes in $\mathbf{U}$ and $\mathbf{V}$. The edge value $\mathbf{e}_{ij}$ is generated by a pairwise function $f(\cdot)$ of the node representations $\mathbf{u}_i$ and $\mathbf{v}_j$. We then remove all the edges whose values are below a threshold τ. We investigate multiple choices of the pairwise function $f(\cdot)$, including the cosine distance, the dot product and the multi-layer perceptron (MLP):

$$f_{cosine}(u_i, v_j) = \frac{\mathbf{u}_i{}^T \mathbf{v}_j}{\|\mathbf{u}_i\| \|\mathbf{v}_j\|}, \tag{1}$$

$$f_{dot}(u_i, v_j) = \mathbf{u}_i^T \mathbf{v}_j, \tag{2}$$

$$f_{MLP}(u_i, v_j) = MLP(\mathbf{u}_i, \mathbf{v}_j). \tag{3}$$

As the node in a set can be connected to multiple nodes on the other side, we want to further find the edges that can reveal the one-to-one region correspondence in two images. We use the maximum bipartite matching to represent such correspondence, where the definitions are as follows:

Definition 3.1. A matching in a bipartite graph is a set of the edges chosen in such a way that no two edges share an endpoint.

Definition 3.2. A maximum matching in a bipartite graph is the matching that contains the maximum number of edges.

We adopt the Ford-Fulkerson Algorithm[1] to find the maximum bipartite matching in our implementation. The found maximum matching enables us to establish the best matching flows between the corresponded regions in two images, which are assumed as the co-occurrent regions. Finally, we re-weight all the node representations in the matching positions. This can be done by multiplying the original feature representations with a mask where the values in the matching positions are α while others are 1. Before directly multiplying the mask with the features, we apply the Gaussian filter to the mask to make it smooth. The enhanced representations are then sent to the rest network components to generate the CAMs. We also generate CAMs prediction based on the features before MBM module, then we use the enhanced CAMs to supervise the learning of this prediction, which serves as an auxiliary self-supervised loss.

3.2 Intra-Graph Message Passing

As directly classifying each position in the feature map is prone to generate sparse prediction maps, we propose to propagate the feature representations with graphs. We first model the feature representations in an image as an undirected grid-like graph where the nodes are the feature points at all locations. Then, each graph node $\mathbf{u}_i$ is updated by attending all neighboring nodes $u_j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is all neighboring nodes of node $\mathbf{u}_i$ and selectively aggregating information from them. To that end, we first generate a relevance score e_{ij} between the center node $\mathbf{u}_i$ and the neighboring node $\mathbf{u}_j$ by computing the dot product of them in a learnable space:

$$e_{ij} = \theta(\mathbf{u}_i)^T \phi(\mathbf{u}_j). \tag{4}$$

Then we the normalize the relevance scores with the softmax function, which serves as the weights when aggregating the representations of neighboring nodes:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{j\in\mathcal{N}_i} \exp(e_{ij})}, \tag{5}$$

$$\mathbf{u}_i^{agg} = \sum_{j\in\mathcal{N}_i} a_{ij} \psi(\mathbf{u}_j). \tag{6}$$

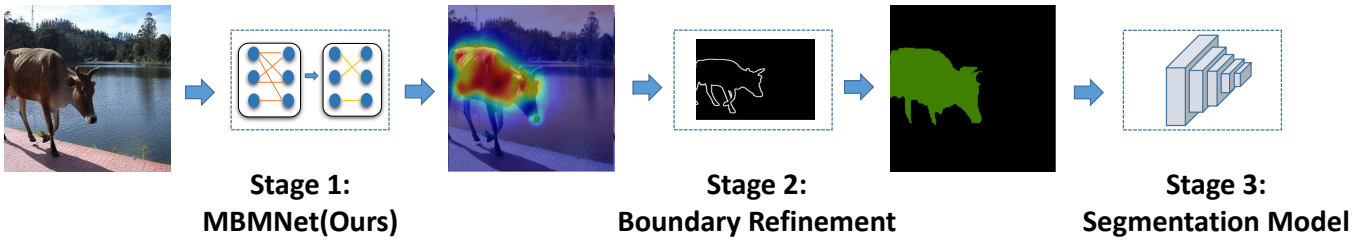

Figure 1: The pipeline of our framework for weakly-supervised segmentation. The whole framework contains three stages. This paper focuses on the first stage which generates the CAMs of input images, which indicate confident seed object regions. In the second stage, we adopt the boundary refinement method proposed in [2] to refine object boundaries. In the third stage, we use the pseudo ground truth masks from the second stage to train a segmentation model in a fully supervised manner.

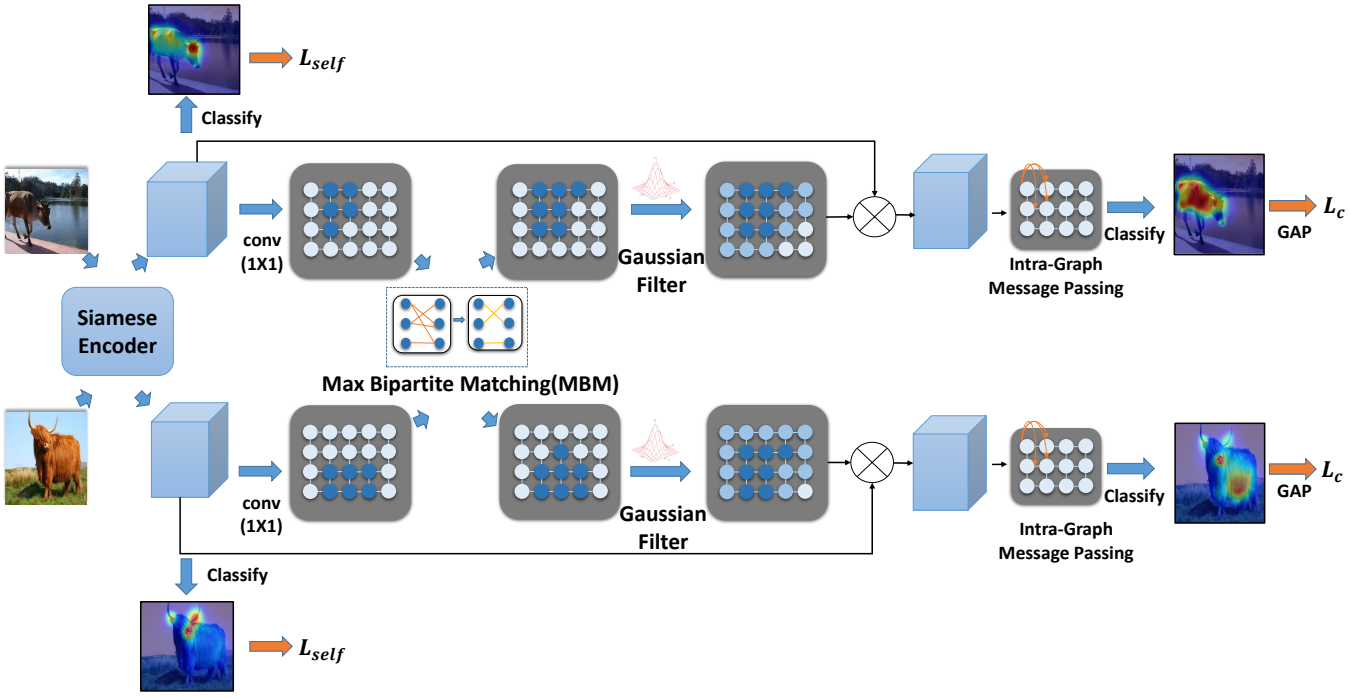

Figure 2: Overall architecture of our methods. Given a pair of sampled images that contain common classes as the network input, we first encoder them into feature representations with a parameter-shared Siamese encoder, then we reinforce the representations on the co-occurrent object regions, which are located by the Maximum Bipartite Matching Module. We further refine the representations in individual images by propagating information between neighboring regions with a graph. Besides the standard multi-label classification loss $\mathcal{L}_c$, we also add an auxiliary loss $\mathcal{L}_{self}$ that uses the refined prediction to supervise the training of the original prediction.

Intuitively, the value of a_{ij} controls how much information will be flowed from node $\mathbf{u}_j$ to $\mathbf{u}_i$. Finally, we fuse the original representations with the aggregated representations by addition:

$$\mathbf{u}'_i = \mathbf{u}_i + \mathbf{u}_i^{agg}. \tag{7}$$

In the above equations, $\theta(\cdot)$, $\phi(\cdot)$ and $\psi(\cdot)$ are linear transformations followed by the ReLU non-linearity. Particularly, $\theta(\cdot)$ and $\phi(\cdot)$ construct a space for computing node similarity and $\psi(\cdot)$ encodes the information for propagation. We investigate multiple types of graphs to represent an image in our experiments. The main difference lies in the adjacent matrix which defines the neighborhood of a graph node. As image pixels have the 2-dimensional grid structure, we can determine the node connections based the pixel connectivity. Here, we experiment with three choices of pixel connectivity that are widely used for images: 1) Von Neumann neighborhood. Given a pixel coordinate (x, y), the Von Neumann neighborhood refers to the pixels that are horizontally and vertically connected, *i.e.* $(x \pm 1, y)$ or $(x, y \pm 1)$. There are at most 4 neighboring nodes in this case. 2) Moore neighborhood. Compared with the Von Neumann neighborhood, the Moore neighborhood also includes the diagonally connected pixels, *i.e.* $(x \pm 1, y \pm 1)$. There are at most

8 neighboring pixels in this case. 3) Full connection. An image is modeled with a fully connected graph, where all pixels are connected with each other. Figure.3 illustrates the three types of pixel connectivity.

3.3 Training

In the training process, the network is trained by sampling a pair of images that contain at least one common class. Our training objective includes a multi-label soft cross-entropy classification loss $\mathcal{L}_{class}$ which is applied on the CAMs after global average pooling, and a self-supervised loss $\mathcal{L}_{self}$ that computes the mean squared error between CAMs before and after refinement:

$$\mathcal{L}_{class} = -\sum_{i=1}^{N}(y[i] * log(\frac{1}{1+exp(-x[i])}) + (1-y[i]) * log(exp(\frac{-x[i]}{1+exp(-x[i])}))), \tag{8}$$

$$\mathcal{L}_{self} = \sum_{i=1}^{HW}(\mathbf{m'}_i - \mathbf{m}_i)^2, \tag{9}$$

where x denote the predicted probability of class i and $y[i]$ denote the ground truth label of the i_{th} class. m and m' denote the predicted probability vectors in the CAMs before and after enhancement. We balance the two losses with hyper-parameter λ:

$$\mathcal{L} = \mathcal{L}_{class} + \lambda\mathcal{L}_{self}. \tag{10}$$

4 EXPERIMENT

4.1 Dataset and Evaluation Metric

Our network is trained and evaluated on the PASCAL VOC 2012 dataset [10]. Aligned with the experiment set-ups in previous works, the training set is the original training images plus images from [14]. Finally, there are 10,582 images for training, 1,449 images for validation and 1,456 for testing. We use the standard mean Intersection-over-Union(mIoU) as the evaluation metric for all experiments.

4.2 Implementation Details

We employ ResNet-50 as the network backbone of the proposed MBMNet, which is pre-trained on ImageNet [20]. The best performance of our network is achieved when setting the enhancement scalar α as 1.3, the threshold τ in the MBM as 0.35 and λ in Equation 10 as 10, and using Von Neumann connectivity for message passing in the graph. If not specified, we use them as the default parameters in our experiments. Our network is built upon the PyTorch library. We use SGD as the optimizer with a mini-batch of 32 images and train the network for 5 epochs. The learning rate is initially set to 0.02, and decreases at every iteration with polynomial decay [36] of 0.9. The weight decay set to 0.004. We apply random crop, random scale, and random flip to the images as data augmentation in the training process.

After the network training in the first stage, we generate the CAMs of all training images by sampling a pair of images that contain at least one same class and send them to the network. In the third stage, we train a DeepLab-LargeFOV (Resnet-50 based) [7] as the segmentation model with the generated pseudo ground truth masks. During the inference phase, we adopt the multi-scale input testing that re-scales the input images to [0.5, 1, 1.5,2] and fuse their predictions by average. We use denseCRF [28] as a post-processing stage to further refine the result.

MBM	Intra-Graph	Self-supervision	mIoU(%)
			48.3
✓			49.5
	✓		49.3
✓	✓		49.8
✓	✓	✓	**50.2**

Table 1: Ablation experiments on different modules in our method. MBM denotes the use of the Maximum Bipartite Matching Module, Intra-Graph denotes the message propagation mechanism, and Self-supervision denotes the auxiliary self-supervised loss. Every module brings performance improvement over the baseline model.

4.3 Ablative Analysis

The goal of this session is to inspect the effectiveness of each component in our network. All the experiments in this section are conducted on PASCAL VOC 2012 [10] training set, and we report the performance with standard mIoU score. Our compared baseline is the model that removes the two proposed modules and directly generates the CAMs of each image. Then we gradually add our proposed modules to the baseline network for comparison. The results are shown in Table. 1. As we can see, both of our proposed modules can boost performance effectively. When they are combined along with the self-supervised loss, our network can achieve the best performance. In Figure 4, we present some visualization examples that compare the results of our method with the results of the baselines. Our method can effectively expand object regions that are closer to the ground truth object regions. Next, we investigate the variants of the network designs and the choices of the hyper-parameters in the network.

Pairwise function $f(\cdot)$. As is discussed in Section. 3.1, the pairwise function $f(\cdot)$ has multiple choices, including the cosine distance, the dot product and a two-layer MLP that predicts a value. We compare different choices in Table. 2. We add a sigmoid function after $f_{dot}(\cdot)$ and $f_{MLP}(\cdot)$ such that all functions output the value in the range of [0,1]. We also test different threshold values τ that determine the graph connections, as described in Section. 3.1. We choose the values from 0.25 to 0.95 with an interval of 0.1. The results in Table. 2 show that the pairwise function that computes the dot product yields the best performance. The choice of τ also has a clear influence on the performance. When using cosine as the pairwise function $f(\cdot)$, the best performance is achieved when setting the threshold value τ as 0.35.

MBM vs. Attention. To determine the common object regions with two images, we can also achieve this goal by attention mechanism. We create two baseline methods that adopt the attention mechanism for comparison. Concretely, the attention model first generates the similarity values between all pixels in two images. Then, to determine the confidence of predicting a pixel location

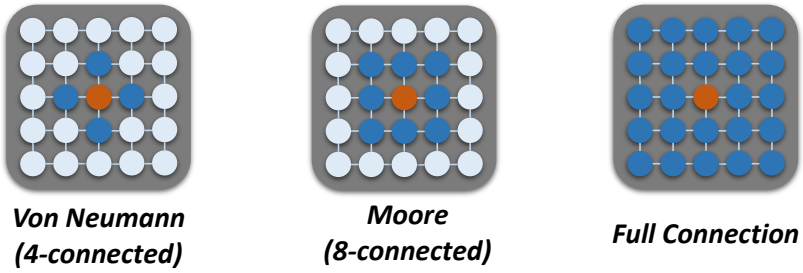

Figure 3: The three types of pixel connectivity. The orange node denotes the center, the blue nodes denote the neighboring nodes.

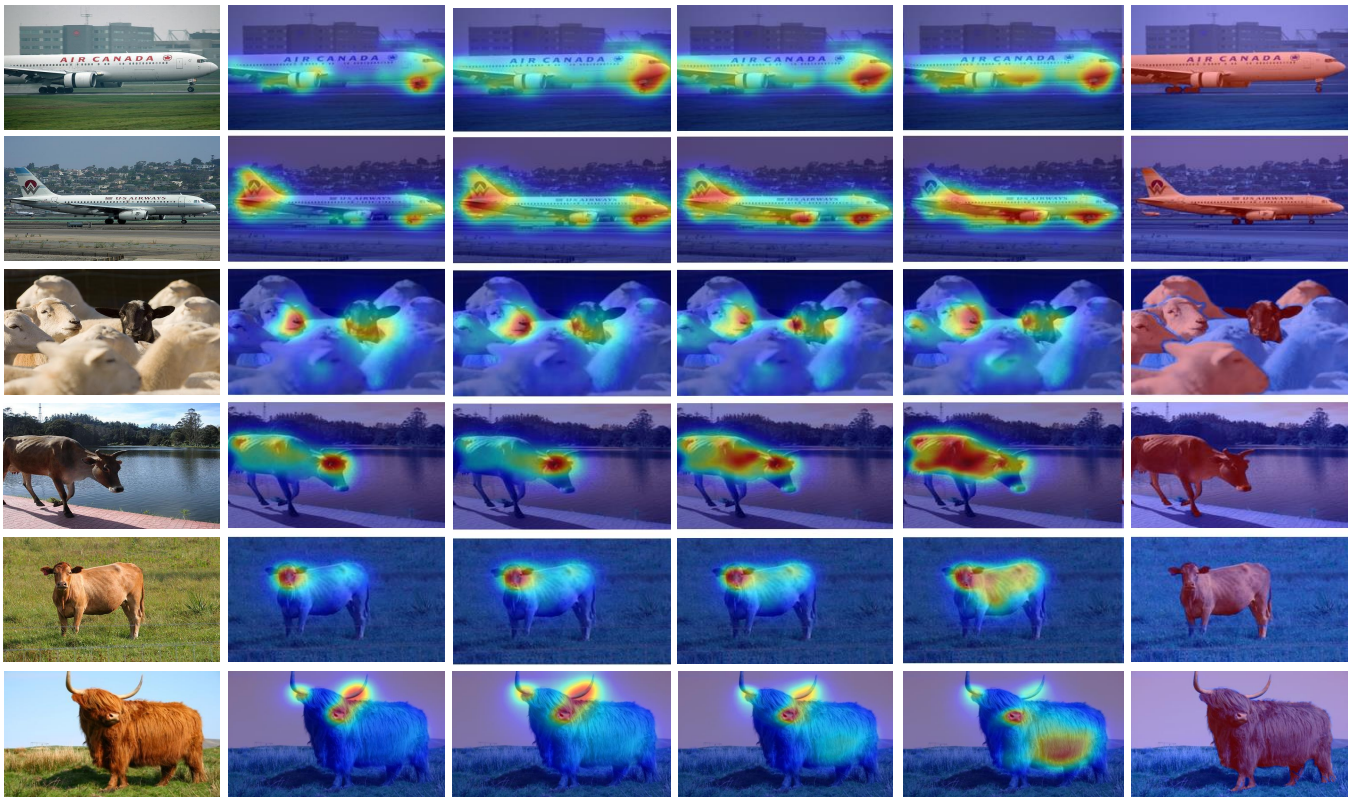

Figure 4: Visualization examples that compare the CAMs generated by our method with the CAMs generated by baseline methods. We gradually add our proposed modules to the baseline model for comparison. Our method can effectively expand object regions that are closer to the ground truth object regions.

as the foreground, we can take the maximum values among all similarity scores corresponding to this pixel, or we can sum all similarity scores corresponding to this pixel. We use dot product to compute the similarity value in the attention model, as is done in our MBM module, and then use the above two methods to generate the confidence of each location followed by a sigmoid layer. Finally, the confidence map is multiplied with the original features to re-weight the representations. We denote these two baseline models by *attention-max* and *attention-sum*. We compare these methods in Table. 4. As we can see, our proposed MBM module achieves better performance than the attention-based method, as the matching algorithm in our method can find the one-to-one correspondence with high confidence. For the attention models, using the maximum value among all connections to compute a confidence map is slightly better than using the summed value.

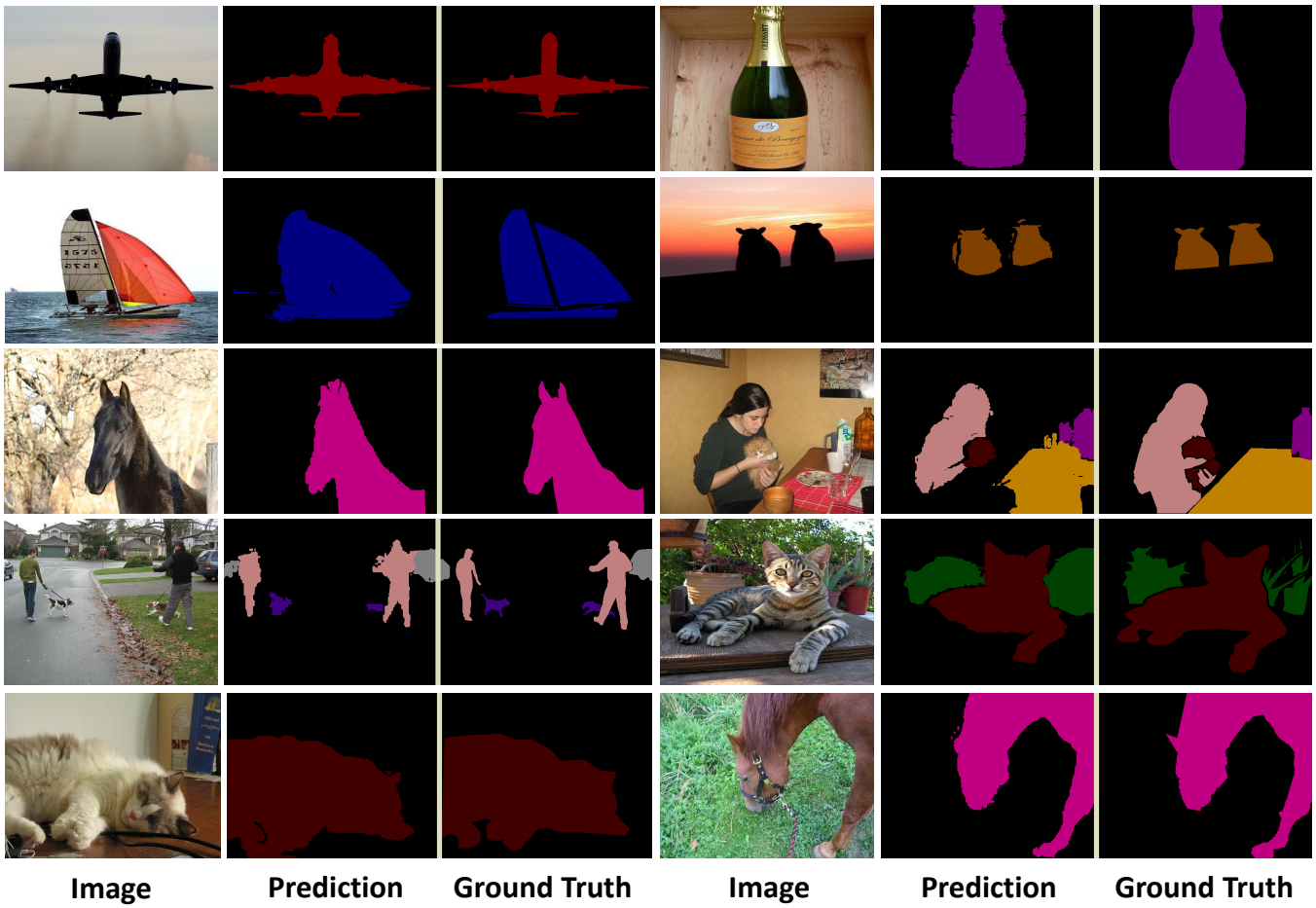

Figure 5: Qualitative results with examples on the PASCAL VOC dataset. Our algorithm generates high-quality prediction in the weakly supervised segmentation task.

Threshold	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Dot product	47.3	49.6	49.4	50.0	49.6	49.8	49.6	49.7
Consine	49.8	**50.2**	49.4	48.8	48.1	47.5	47.3	46.0
MLP	50.0	49.4	49.3	48.3	49.8	49.7	50.0	49.3

Table 2: Comparison different threshold with different pairwise function $f(\cdot)$, results are the pseudo label in VOC 2012 train set and reported in mIoU(%) , the dot product function with threshold 0.55 achieve the best performance.

Connect Type	No. Pixels Connected	mIoU(%)
Von Neumann	4	**50.2**
Moore	8	49.8
Full connection	Many	49.9

Table 3: Comparison with different graph connectivity in the intra-graph module. The Graph constructed with Von Neumann node connectivity achieves the best performance.

Methods	mIoU(%)
Attention-Max	49.5
Attention-Sum	49.3
Our-MBM	**50.2**

Table 4: Comparison of our MBM with attention-based model variants. Our graph-matching based method achieves better performance than the attention-based baselines.

Graph connectivity. In Table. 3, we compare different ways to define graph neighbourhood as motioned in Section. 3.2. We find that all three kinds of neighborhood work well in the intra-graph message passing module. Particularly, the Von Neumann neighborhood (4 connected nodes) yields the best performance. The possible reason is that since the Von Neumann only considers the four nearest nodes, the information exchanging is more confident and reliable.

methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	**mIoU**
val set (Ours)	88.9	81.0	33.7	84.8	65.0	66.8	85.9	79.0	75.0	32.3	78.5	34.6	75.9	79.6	73.7	58.6	47.0	82.7	36.5	67.2	64.4	66.2
test set (Ours)	89.9	85.0	33.4	82.2	59.2	67.7	86.2	83.0	74.4	29.4	80.6	42.4	78.0	80.9	79.6	61.7	43.5	86.4	46.4	58.0	61.1	67.1

Table 5: The detail results on PASCAL VOC 2012 *val and test set.*

Method	AS	Val	Test
FCN-MIL [42]	-	25.7	24.9
CCNN [41]	-	35.3	35.6
EM-Adapt [40]	-	38.2	39.6
DCSM [47]	-	44.1	45.1
BFBP [45]	-	46.6	48.0
SEC [27]	-	50.7	51.7
CBTS [44]	-	52.8	53.7
TPL [26]	-	53.1	53.8
MEFF [12]	-	-	55.6
PSA [3]	-	61.7	63.7
IRN [2]	-	63.5	64.8
SSDD. [48]	-	64.9	65.5
MBMNet (w/o crf)	-	64.8	65.6
MBMNet (w/ crf)	-	**66.2**	**67.1**

Table 6: Comparison of the weakly supervised semantic segmentation methods. With the setting of without any additional supervision, our method outperforms all the previous methods on both validation set and test set.

Method	AS	Val	Test
MCNN [50]	WV	38.1	39.8
AFF [43]	S	54.3	55.5
STC [56]	S + WI	49.8	51.2
AE-PSL [55]	S	55.0	55.7
WebS-i2 [22]	WI	53.4	55.3
MDC [57]	S	60.4	60.8
MCOF [54]	S	60.3	61.2
DSRG [18]	S	61.4	63.2
AISI [11]	IS	63.6	64.5
FickleNet [29]	S	64.9	65.3
DSRG+EP. [52]	S	61.5	62.7
OAA+. [21]	S	65.2	66.4
MBMNet (w/o crf)	-	64.8	65.6
MBMNet (w/ crf)	-	**66.2**	**67.1**

Table 7: Comparison of the weakly supervised semantic segmentation methods with more additional supervision, our method outperforms all the previous methods on both validation set and test set even we do not use any additional supervision. WV denote Web Video, S denote Saliency Mask, WI denote Web Image, IS denote Instance Image.

CAMs	CAMs + BR	Enhanced CAMs	Enhanced CAMs + BR
48.3	66	50.2	66.8

Table 8: The performance of the synthesized segmentation labels in mIoU, evaluated on the PASCAL VOC 2012 training set. BR [2]: Boundary Refinement

4.4 Analysis of Pseudo Labels

The performance of our synthesized segmentation labels is measure with standard mIoU on PASCAL VOC 2012 [10] training data. As shown in the Table 8, with the same boundary refinement methods [2], the synthesized segmentation labels with our MBMNet enhanced CAMs performance better than the CAMs without enhancement which consistent with our initial statement in Section 1 that the better CAMs always generate better pseudo ground truth.

4.5 Comparison with the State-of-the-art Results

Finally, we compare our network with state-of-the-art methods on the PASCAL VOC 2012 dataset including both the validation set and the testing set. Table. 6 shows the performance of different methods under the setting that only image labels are available for weakly supervised segmentation. We achieve the state-of-the-art performance on both the validation set and the testing set. We also compare our method with previous works that employ additional information as supervision, and the results are shown in Table. 7. Although these methods adopt various additional information, such as saliency masks[5, 17, 18, 30], instance saliency map [11] and web images [46], our method still outperforms all previous results and achieves new state-of-the-art performance. In Figure. 5, we present some qualitative results on PASCAL VOC 2012.

5 CONCLUSION

In this paper, we adopt graph-based methods to improve the CAMs generation network, which is a crucial step in the weakly supervised segmentation model. To discover more object regions in the CAMs, we send a pair of images that contain common object categories to the network, and use the maximum bipartite matching(MBM) algorithm to find matching areas in two images which are used to enhance feature representations. To further refine the feature maps of images, we propose to propagate feature information between neighboring regions. Experiment results on the Pascal VOC dataset validate our contributions and we achieve new state-of-the-art performance on both validation set and test set.

ACKNOWLEDGEMENTS

This work is supported by the Delta-NTU Corporate Lab with funding support from Delta Electronics Inc. and the National Research Foundation (NRF) Singapore (SMA-RP10). This work is also partly supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-RP-2018-003) and the MOE Tier-1 research grants: RG126/17 (S), RG28/18 (S) and RG22/19 (S).

REFERENCES

[1] 2020. Ford–Fulkerson algorithm. *Wikipedia* (Feb 2020). https://en.wikipedia.org/wiki/Ford\T1\textendashFulkerson-algorithm

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. 2019. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. In *CVPR*.

[3] Jiwoon Ahn and Suha Kwak. 2018. Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation. In *CVPR*.

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

[5] Arslan Chaudhry, K. Puneet Dokania, and H.S. Philip Torr. 2017. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *Proc. of British Machine Vision Conference*.

[6] Hong Chen, Yifei Huang, and Hideki Nakayama. 2018. Semantic aware attention based deep object co-segmentation. In *Asian Conference on Computer Vision*. Springer, 435–450.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

[8] Jifeng Dai, Kaiming He, and Jian Sun. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1635–1643.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111, 1 (2015), 98–136.

[11] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, R. Ralph Martin, and Shi-Min Hu. 2018. Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation. In *ECCV*.

[12] Weifeng Ge, Sibei Yang, and Yizhou Yu. 2018. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*.

[13] Zaïd Harchaoui and Francis Bach. 2007. Image classification with segmentation graph kernels. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. 2011. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*. IEEE, 991–998.

[15] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[16] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. 2017. Weakly Supervised Semantic Segmentation using Web-Crawled Videos. In *CVPR*.

[17] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. 2018. Self-Erasing Network for Integral Object Attention. In *NIPS*.

[18] Zilong Huang, Wang Xinggang, Wang Jiasi, Wenyu Liu, and Wang Jingdong. 2018. Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing. In *CVPR*.

[19] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. 2019. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 869–878.

[20] D. Jia, B. Alex, S. Sanjeev, S. Hao, K. Aditya, and L. Fei-Fei. 2012. IMAGENET Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). http://www.image-net.org/challenges/LSVRC/2012/index.

[21] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. 2019. Integral Object Mining via Online Attention Accumulation. In *ICCV*.

[22] Bin Jin, Maria V. Ortiz Segovia, and Sabine Susstrunk. 2018. Webly Supervised Semantic Segmentation. In *CVPR*.

[23] Armand Joulin, Francis Bach, and Jean Ponce. 2012. Multi-class cosegmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 542–549.

[24] Dagmar Kainmueller, Florian Jug, Carsten Rother, and Gene Myers. 2014. Active graph matching for automatic joint segmentation and annotation of C. elegans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 81–88.

[25] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 876–885.

[26] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. 2017. Two-phase learning for weakly supervised object localization. In *ICCV*.

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

[28] Philipp Krähenbühl and Vladlen Koltun. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*. 109–117.

[29] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. 2019. FickleNet: Weakly and Semi-supervised Semantic Image Segmentation. In *CVPR*.

[30] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernest, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *CVPR*.

[31] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. 2018. Deep object co-segmentation. In *Asian Conference on Computer Vision*. Springer, 638–653.

[32] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3159–3167.

[33] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3159–3167.

[34] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1925–1934.

[35] Weide Liu, Guosheng Lin, Tianyi Zhang, and Zichuan Liu. 2020. Guided Co-Segmentation Network for Fast Video Object Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[36] Wei Liu, Andrew Rabinovich, and Alexander C Berg. 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579* (2015).

[37] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. 2020. CRNet: Cross-Reference Networks for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4165–4173.

[38] Charles Iury Oliveira Martins, Roberto Marcondes Cesar, Leonardo Ré Jorge, and André Victor Lucci Freitas. 2011. Segmentation of similar images using graph matching and community detection. In *International Workshop on Graph-Based Representations in Pattern Recognition*. Springer, 265–274.

[39] G Papandreou, L-Ch Chen, K Murphy, and AL Yuille. [n.d.]. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. arXiv, 2015. *arXiv preprint arXiv:1502.02734* ([n. d.]).

[40] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille. 2015. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*.

[41] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. 2015. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*.

[42] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2015. Fully convolutional multi-class multiple instance learning. In *ICLR*.

[43] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. 2016. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*.

[44] Anirban Roy and Sinisa Todorovic. 2017. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*.

[45] Fatemehsadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M. Alvares. 2016. Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation. In *ECCV*.

[46] Tong Shen, Guosheng Lin, Chunhua Shen, and Reid Ian. 2018. Bootstrapping the Performance of Webly Supervised Semantic Segmentation. In *CVPR*.

[47] Wataru Shimoda and Keiji Yanai. 2016. Distinct Class Saliency Maps for Weakly Supervised Semantic Segmentation. In *ECCV*.

[48] Wataru Shimoda and Keiji Yanai. 2019. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 5208–5217.

[49] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. 2020. Conditional Gaussian Distribution Learning for Open Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13480–13489.

[50] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. 2016. Weakly-Supervised Semantic Segmentation using Motion Cues. In *ECCV*.

[51] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. 2011. Object cosegmentation. In *CVPR 2011*. IEEE, 2217–2224.

[52] Weitao Wan, Jiansheng Chen, Tianpeng Li, Yiqing Huang, Jingqi Tian, Cheng Yu, and Youze Xue. 2019. Information Entropy Based Feature Pooling for Convolutional Neural Networks. In *ICCV*.

[53] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. 2014. Learning Fine-grained Image Similarity with Deep Ranking. In *CVPR*. 1386–1393.

[54] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. 2018. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*.

[55] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *CVPR*.

[56] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. 2017. STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation. In *IEEE Trans. on PAMI*.

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

[58] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z Li. 2019. Unsupervised Graph Association for Person Re-Identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 8321–8330.

[59] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*. 1395–1403.

[60] Junchi Yan, Minsu Cho, Hongyuan Zha, Xiaokang Yang, and Stephen M Chu. 2015. Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE transactions on pattern analysis and machine intelligence* 38, 6 (2015), 1228–1242.

[61] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. 2017. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE international conference on computer vision*. 5142–5150.

[62] F. Yu and V. Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*.

[63] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. DeepEMD: Few-Shot Image Classification with Differentiable Earth Mover's Distance and Structured Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12203–12213.

[64] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. 2019. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 9587–9595.

[65] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. 2019. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5217–5226.

[66] T. Zhang, G. Lin, J. Cai, T. Shen, C. Shen, and A. Kot. 2019. Decoupled spatial neural attention for weakly supervised semantic segmentation. *IEEE Transactions on Multimedia* 21, 11 (2019), 2930–2941.

[67] T. Zhang, G. Lin, W. Liu, J. Cai, and A. Kot. 2020. Splitting vs. merging: mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *European Conference on Computer Vision*.

[68] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.