Treating Pseudo-labels Generation as Image Matting for Weakly Supervised Semantic Segmentation

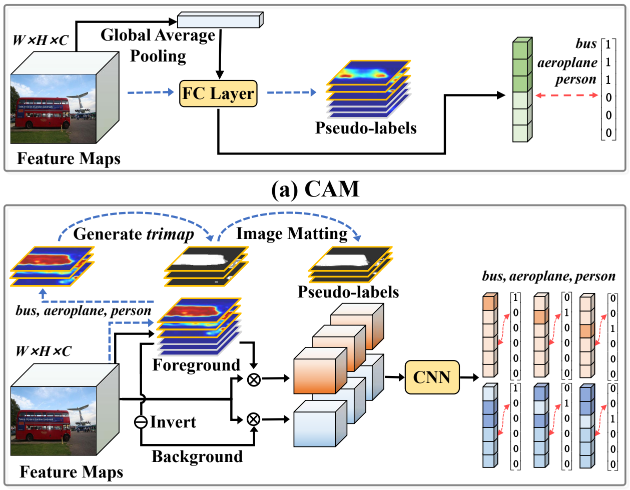
Changwei Wang[1,3,*], Rongtao Xu[1,3,*], Shibiao Xu[2,†], Weiliang Meng[1,3,†], Xiaopeng Zhang[1,3]
[1]The State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
shibiaoxu@bupt.edu.cn, weiliang.meng@ia.ac.cn

Abstract

Generating accurate pseudo-labels under the supervision of image categories is a crucial step in Weakly Supervised Semantic Segmentation (WSSS). In this work, we propose a Mat-Label pipeline that provides a fresh way to treat WSSS pseudo-labels generation as an image matting task. By taking a trimap as input which specifies the foreground, background and unknown regions, the image matting task outputs an object mask with fine edges. The intuition behind our Mat-Label is that generating trimap is much easier than generating pseudo-labels directly under weakly supervised setting. Although current CAM-based methods are off-the-shelf solutions for generating a trimap, they suffer from cross-category and foreground-background pixel prediction confusion. To solve this problem, we develop a Double Decoupled Class Activation Map (D2CAM) for Mat-Label to generate a high-quality trimap. By drawing on the idea of metric learning, we explicitly model class activation map with category decoupling and foreground-background decoupling. We also design two simple yet effective refinement constraints for D2CAM to stabilize optimization and eliminate non-exclusive activation. Extensive experiments validate that our Mat-Label achieves substantial and consistent performance gains compared to current state-of-the-art WSSS approaches.

1. Introduction

Semantic segmentation is a fundamental task in computer vision that aims to assign pixel-level semantic labels to objects in an image. In the last decade, the boom in deep neural networks has facilitated the rapid development of deep learning based semantic segmentation methods [3].



(a) CAM

(b) Mat-Label (ours)

Figure 1. The schematic diagram of the pseudo-labels generation pipeline based on the common CAM and our Mat-Label. The **black line** is the training phase workflow, and the **blue line** is the inference phase workflow, and the **red line** is the loss calculation.

However, fully supervised semantic segmentation models rely on expensive and time-consuming pixel-level annotations. Therefore, weakly supervised semantic segmentation (WSSS) has attracted increasing attention, as it is trained only by inexpensive annotations (*e.g.* scribbles [31], bounding box [9], or image-level [47] labels). In this work, we focus on WSSS based on image-level labels, which aims to use only object category labeling supervision and is considered to be the cheapest yet most challenging setting.

Most of existing image-level WSSS methods usually adopt a pipeline containing the following three sequential steps: ***i)*** generating pseudo-labels through class activation maps, *e.g.* CAM [47]; ***ii)*** refining the pseudo-labels as segmentation ground truth masks; ***iii)*** training the segmenta-

* Equal Contribution and † Corresponding Authors.

tion network under pseudo ground truth masks supervision. Since the quality of the generated pseudo-labels in the first step has a great impact on the final segmentation results, many CAM-based variants [47, 40, 23] have tried to generate better pseudo-labels under the supervision of only image category labels. However, the generation of complete and edge-accurate pseudo-labels is still unsatisfactory and challenging for existing CAM-based methods.

As a technique ignored by the current WSSS approaches, image matting [29] aims to utilize the input priori information (*i.e.* *trimap* that specifies the foreground regions, background regions, and unknown regions) to predict the probability that each pixel belongs to the foreground object (*i.e.* alpha matte α). There are two reasons that motivate us to explore the use of image matting to generate WSSS pseudo-labels. **First**, the alpha matte derived from image matting gives high response to foreground objects and low response to backgrounds, so the region of the object can be easily determined. In addition, alpha matte has more complete object activation and accurate edge representation than common CAM [47], as shown in Figure 1 and Figure 2. With the above attractive properties, alpha matte can be used as an available form of pseudo-labels. **Second**, it is difficult to obtain complete and accurate activation of objects through CAM [47] by relying on image category labels alone, but it is much easier to construct the *trimap* priori required for image matting. Specifically, *trimap* only needs to distinguish the foreground from the background as much as possible and marks the uncertain regions as unknown.

Motivated by the above discussions, we propose a pseudo-labels generation pipeline called Mat-Label, which treats WSSS pseudo-labels generation as an image matting task, as shown in Figure 1 (b). Since the quality of *trimap* has a great impact on the final alpha matte [29], how to obtain a high-quality *trimap* with good foreground-background division under image category supervision only becomes the most urgent challenge for our Mat-Label to solve. The off-the-shelf solution is to use the classical CAM [47] or its variants to build the *trimap*. However, there are two common flaws often exist with these CAM-based works: ***1)*** The foreground region is incompletely activated, or the background region is over-activated; ***2)*** Some foreground object pixels are classified in the wrong category. The ***flaw 1)*** is caused by the lack of unawareness and modeling of foreground and background in existing CAM-based methods, while the ***flaw 2)*** is due to the non-exclusive activation [7] of different classes during the training of the multi-label classifier. These two flaws prevent existing CAM-based methods from constructing a *trimap* with good foreground-background division.

To alleviate the above two flaws, we can generate class activation maps by explicitly modeling the foreground and background, inspired by recent approaches [35, 27, 41, 42] to model background in weakly supervised learning. However, these previous methods are designed to model category-independent background regions (*e.g.* sky, grass, and sea) and are not competent to category-dependent WSSS pseudo-labels generation task. Therefore, we develop a Double Decoupled Class Activation Map (D2CAM) to construct a high-quality *trimap* for our Mat-Label pipeline, in which we design a new class activation map learning strategy with both category decoupling and foreground-background decoupling to accommodate the WSSS task. On the one hand, we set category-separated classification ground truth to optimize the class activation map for each category individually, as shown in Figure 1 (b). On the other hand, we design a special foreground-background decoupled learning that is different from previous approaches, drawing on the idea of triplet loss [10] in metric learning. Specifically, we encourage the classification vector obtained from the foreground region to be close to the ground truth, while the classification vector from the background region is far from the ground truth to obtain a better foreground-background division and thus alleviate the ***flaw 1)***. In addition, we design two simple yet effective class activation map refinement constraints to stabilize the optimization process and mitigate the ***flaw 2)*** by suppress non-exclusive activation, respectively. **The main contributions of this work are summarized as follows:**

- We propose an interesting and promising WSSS pseudo-labels generation pipeline called Mat-Label, which is the first solid baseline to treat WSSS pseudo-labels generation as an image matting task;
- We develop a Double Decoupled Class Activation Map (D2CAM) to generate category-dependent *trimap* with good foreground-background division for Mat-Label;
- We explicitly model class activation map with category decoupling and foreground-background decoupling by constructing category-separated classification ground truth and drawing on the idea of metric learning;
- We design two simple but effective refinement constraints for class activation map training to stabilize optimization and eliminate non-exclusive activation.

Extensive experiments validates that our Mat-Label pipeline substantially achieves the state-of-the-art performance of WSSS on both PASCAL VOC 2012 and MS COCO 2014 datasets.

2. Related Work

Pseudo-labels Generation for WSSS. Extracting a CAM [47] has been the standard step for generating pseudo-initial masks in WSSS. However, focuses on the most discriminative parts of the object, resulting in an incomplete mask since it's trained for classification. Thus, several CAM-based variants have been developed to address this issue. Erasure-based approaches [14, 20, 46, 44] use iterative erasure strategies that retain erased features to prevent the classification network from focusing only on discriminative object parts. SEC [18] proposes three principles, *i.e.*, seed, expand, and constrain, to refine CAMs, which are followed by many other works. Some methods [40, 23] aggregate different contexts by introducing feature maps from different locations or layers to obtain a larger activation region. ReCAM [7] reactivates the converged CAM with binary cross-entropy loss using softmax cross-entropy loss. L2G [17] gives an online local-to-global knowledge transfer framework for high-quality object attention mining. More recently, ESOL [30] provide an extension and contraction scheme for deformable convolution-based offset learning to obtain a more complete class activation map. In this work, we propose a new image matting based pseudo-labels generation pipeline that clearly differs from the above methods.

Pseudo-labels Refinement for WSSS. The generated pseudo-labels are usually further enhanced by refinement steps [15, 1, 16, 24]. Despite the rapid development of refinement labeling methods, they still rely on good initial labels. If the initial mask prediction yields incorrect activation regions, applying the refinement step will cover more inaccurate regions. Experimental results show that our Mat-Label cascades with these refinement steps excite more exciting performance, due to the better initialization provided.

Image Matting. The image matting algorithms [29] employ the user input as an additional priori to precisely separate a foreground object for image editing and compositing purposes. The *trimap* is the most common priori information, which provides information about the foreground, background, and unknown regions. Image matting algorithms use the priori information provided by *trimap* to infer the probability that a location region belongs to the foreground. Traditional image matting methods [29, 5] and deep learning based image matting methods [34, 36] have been well-developed and applied in the past decade. In this work, we treat WSSS pseudo-labels generation as an image matting task for the first time. What's even more significant is that we reveal the potential of combining WSSS with image matting, which can provide inspiration for future work.

Background Modeling in Weakly Supervised Learning. Background modeling has been shown to improve the performance of weakly supervised learning in other scopes besides WSSS pseudo-labels generation. Background-aware pooling(BAP) [35] is a background-aware pooling strategy for WSSS with bounding box annotations. Specifically, it uses the region outside the ground truth bounding box as the background context for objects in the inner box. Lee *et al.* [27] propose two background-aware losses to suppress the localization scores of background frames in weakly supervised action localization. BAS [41] improves the performance of weakly supervised object localization by suppressing the activation values of background regions. C^{2}AM [42] models class-agnostic background regions for the refinement of WSSS labels by introducing contrastive learning. Note that C^{2}AM [42] is used to refine the WSSS labels rather than generate the pseudo-labels, so it is not the direct competitor of our work but can be used together. **In general, our D2CAM for generating *trimap* in Mat-Label has the following significant differences compared to the above methods: *1)*** Existing methods aim to model category-independent background, *i.e.* the regions (*e.g.* sky, grass, and sea) that do not contain any category of objects. In contrast, our D2CAM models category-dependent background, *e.g.* for "person" category, "bicycle" around the "person" should also be considered as background regions, so the problem we face is more challenging; ***2)*** Our D2CAM introduces the concept of triplet loss [10] for the first time to model both foreground and background regions in a unified form; ***3)*** Our D2CAM designs two refinement constraints that can further improve the quality of foreground-background modeling; ***4)*** Existing methods are developed for other weakly supervised tasks and cannot be directly applied to pseudo-labels generation for WSSS. To the best of our knowledge, our D2CAM is the first WSSS method to construct class activation maps by foreground-background modeling, while C^{2}AM [42] can only yield category-independent background maps for pseudo-labels refinement and can not competently generate category-dependent class activation maps.

3. Methodology

In this work, we propose a pseudo-labels generation pipeline named Mat-Label to obtain high-quality pseudo masks using only image category labels. The core idea of our Mat-Label is to treat the WSSS pseudo-labels generation as an image matting task [29].

3.1. Preliminaries

Image matting [29] is one of the fundamental tasks in computer vision and is mainly used to accurately separate foreground objects. A natural RGB image can be represented as a linear combination of foreground $F \in \mathbb{R}^{H \times W \times 3}$ and background $B \in \mathbb{R}^{H \times W \times 3}$ with alpha matte $\alpha \in \mathbb{R}^{H \times W}$ as follows:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i,\ \alpha_i \in [0, 1], \tag{1}$$

where H, W denotes the height and the width of the image respectively, while $i \in [H \times W]$ denotes the pixel index. α can be considered as the probability that a pixel belongs to the foreground object, so it can be naturally used as a pseudo-label. In image matting, estimating the opacity value α only based on the given image I is a highly

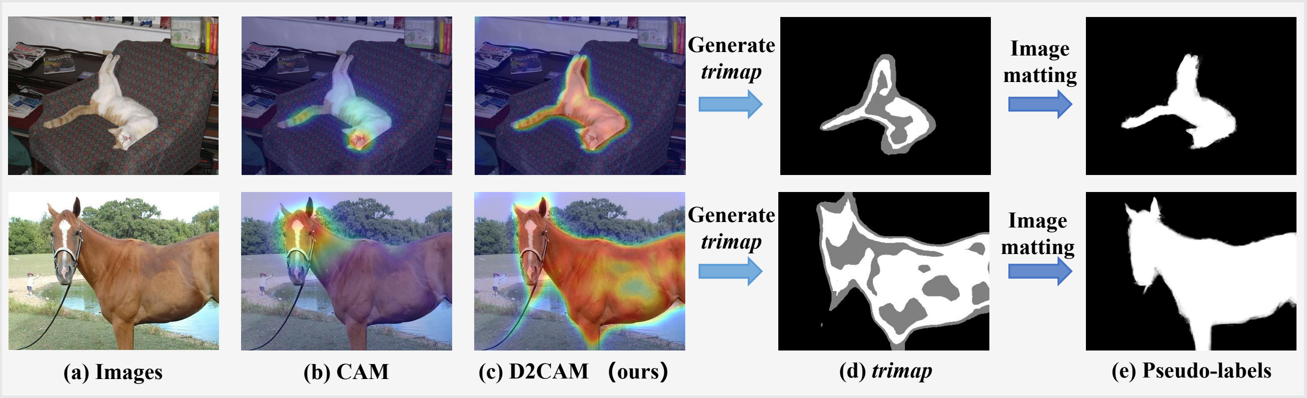


Figure 2. The pipeline of our Mat-Label. **White pixels** in (d) represent the foreground region, **black pixels** in (d) represent the background region, and the **gray pixels** in (d) represent the unknown region. It can be seen that our D2CAM (c) has a more complete and accurate foreground-background division than vanilla CAM [47] (b).

ill-posed problem, and additional priori information (*e.g. trimap*, scribble, background image, *etc*.) is required, in which *trimap* is the most common as it contains the foreground, background, and unknown regions (shown in Figure 2 (d)). Generally speaking, the *trimap* for image matting needs to have the following two properties: First, the *trimap* is **category-dependent**, *i.e.*, each category of objects in one image has its own *trimap*, since image matting can only handle the foreground-background binary classification problem. Second, the *trimap* has an **accurate delineation of foreground and background regions**, which serves as important prior information for performing image matting.

3.2. Mat-Label

Our Mat-Label aims to treat the pseudo-labels generation of WSSS as an image matting task, and Figure 2 illustrates the pipeline of our Mat-Label. In the pipeline, we need to train the network to generate *trimap* with clearly distinguishable foreground and background regions under the supervision of image category labels only. One available option is to generate a *trimap* for each category using the class activation map derived by CAM [47]. Unfortunately, there are foreground over-activations and under-activations in CAM [47] that prevent a good division of foreground and background, as shown in Figure 2 (b).

To automatically construct high-quality *trimap*, we propose a Double Decoupled Class Activation Map (D2CAM) that explicitly models foreground and background regions for each category to alleviate the pixel category confusion and foreground-background confusion problems that exist in CAM-based [47] methods. Instead of adopting the common paradigm of obtaining the class activation map by multiplying the `FC layer` weights with the feature map, we model the class activation map explicitly through well-designed losses. The details of D2CAM will be given in Section 3.2.1. In addition, the process of converting D2CAM's class activation map into *trimap* and eventually generating pseudo-labels will be described in Section 3.2.2.

3.2.1 Generate *trimap* via our D2CAM

Motivation. To obtain class activation map with good foreground and background division to generate high quality *trimap*, we propose Double Decoupled Class Activation Map (D2CAM) that is significantly different from existing CAM-based methods [47]. Specifically, we promote category decoupling and foreground-background decoupling to optimize the class activation map.

Network Architecture. The detailed network architecture of our D2CAM is shown in Figure 3. Generally, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the neural network finally outputs a category prediction vector $\hat{Y} \in \mathbb{R}^{N}$ for the classification. Here H, W are the height and width of the image, while N denotes the number of categories in datasets. Unlike CAM [47] to implicitly optimizes class activation map, we use a lightweight generator $\varphi(\cdot)$ to explicitly predict class activate map, which consists of a single 3×3 convolutional layer and a `sigmoid` activation function layer. Inspired by [49, 41], the classification backbone network is divided into two sub-networks $\mathcal{N}_1$ and $\mathcal{N}_2$ to optimize the class activation map, as shown in Figure 3. The sub-network $\mathcal{N}_2$ includes the last two layers of the classification network, and $\mathcal{N}_1$ includes other previous network layers. First, $\mathcal{N}_1$ extracts the image features by successive convolutional and down-sampling layers to obtain the feature map $\mathcal{F} \in \mathbb{R}^{\frac{1}{8}H\times\frac{1}{8}W\times C}$, with C is the number of channels in the feature map. Specifically, it can be defined as follows:

$$\mathcal{F} = \mathcal{N}_1(I). \tag{2}$$

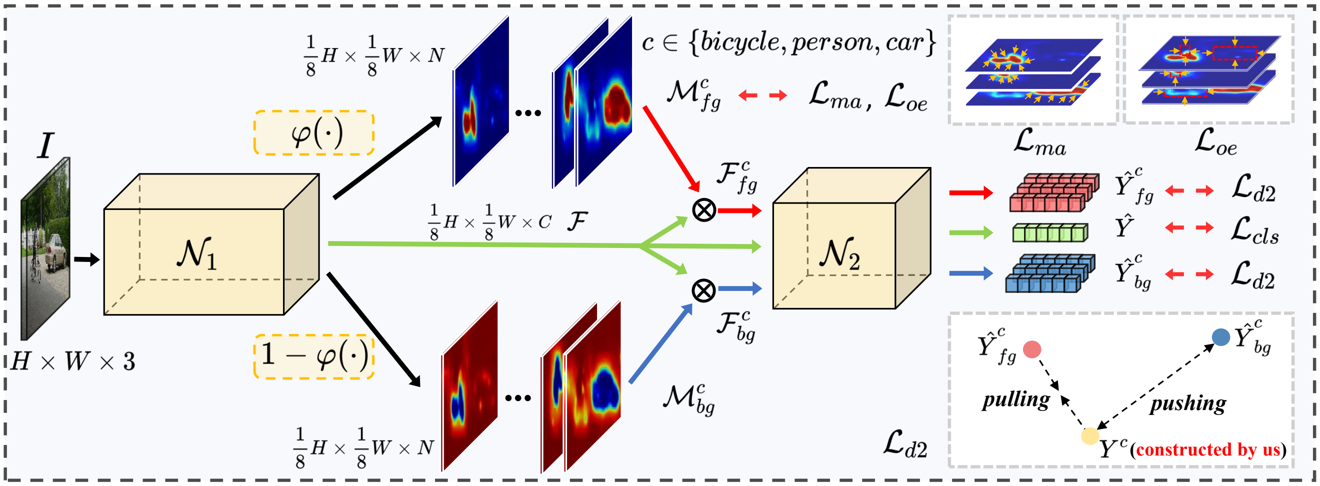

I
$H \times W \times 3$
$\mathcal{N}_1$
$\varphi(\cdot)$
$1-\varphi(\cdot)$
$\frac{1}{8}H \times \frac{1}{8}W \times N$
$\frac{1}{8}H \times \frac{1}{8}W \times C$ $\mathcal{F}$
$\frac{1}{8}H \times \frac{1}{8}W \times N$
$c \in \{bicycle, person, car\}$
$\mathcal{M}^c_{fg}$ $\mathcal{L}_{ma}$, $\mathcal{L}_{oe}$
$\mathcal{F}^c_{fg}$
$\mathcal{F}^c_{bg}$
$\mathcal{M}^c_{bg}$
$\mathcal{N}_2$
$\mathcal{L}_{ma}$
$\mathcal{L}_{oe}$
$\hat{Y}^c_{fg}$ $\mathcal{L}_{d2}$
$\hat{Y}$ $\mathcal{L}_{cls}$
$\hat{Y}^c_{bg}$ $\mathcal{L}_{d2}$
$\hat{Y}^c_{fg}$
$\hat{Y}^c_{bg}$
pulling
pushing
Y^c(constructed by us)
$\mathcal{L}_{d2}$

Figure 3. The details of our D2CAM. We divide the classification network into two parts $\mathcal{N}_1$ and $\mathcal{N}_2$, while the feature map $\mathcal{F}$ is the output of $\mathcal{N}_1$. $\varphi(\cdot)$ is the generator of the class activation map, while $\mathcal{M}_{fg}$ and $\mathcal{M}_{bg}$ refer to the foreground and background class activation maps, respectively. H and W represent the length and width of the input image, while N represents the number of object categories. The white boxes show schematic of the loss functions' principles. Y^c is the category decoupling ground truth constructed by us.

Then, the feature map $\mathcal{F}$ is fed into $\varphi(\cdot)$ to export the foreground class activation map $\mathcal{M}_{fg} \in \mathbb{R}^{\frac{1}{8}H\times\frac{1}{8}W\times N}$, with N is the number of object categories. This means that each channel of $\mathcal{M}_{fg}$ corresponds to a specific category. We further inverse $\mathcal{M}_{fg}$ to obtain the background class activation map $\mathcal{M}_{bg} \in \mathbb{R}^{\frac{1}{8}H\times\frac{1}{8}W\times N}$. The vector $\hat{Y}$ for multi-category classification is obtained by direct input of $\mathcal{F}$ into $\mathcal{N}_2$. Different from existing background modeling methods [49, 35, 41, 42], we promote modeling the category-dependent foreground and background regions separately for each category in the image. Specifically, for a given category of objects, any other category of objects and natural backgrounds (*e.g.* sky, grass, and sea) are all grouped into background regions.

In the training phase, we select specific corresponding channels from $\mathcal{M}_{fg}$ and $\mathcal{M}_{bg}$ for subsequent optimization according to the ground truth of image category labels. Thus, we can obtain category decoupling class activate maps $\mathcal{M}^c_{fg}$ and $\mathcal{M}^c_{bg}$ $c \in \mathcal{C}$, where set $\mathcal{C}$ means the serial number of the category contained in the image. We can then obtain the category decoupled prediction vectors by the following operation:

$$\hat{Y}^c_{fg} = \mathcal{N}_2(\mathcal{F} \otimes \mathcal{M}^c_{fg}), c \in \mathcal{C}; \tag{3}$$

$$\hat{Y}^c_{bg} = \mathcal{N}_2(\mathcal{F} \otimes \mathcal{M}^c_{bg}), c \in \mathcal{C}, \tag{4}$$

where c denotes a specific class of objects contained in the image and $\otimes$ denotes the element-wise product. Here $\hat{Y}^c_{fg}$ and $\hat{Y}^c_{bg}$ contain the context of the category-dependent foreground and background, respectively. Unlike the previous approaches of pure background modeling [49, 35, 41, 42], we need to model both foreground and background regions to achieve category-dependent foreground-background modeling. Here we propose a double decoupled learning that models category-dependent foreground and background regions uniformly, drawing on the idea of triplet loss [10], as shown in $\mathcal{L}_{d2}$ of Figure 3.

Loss Functions. To perform weakly supervised training using only image categories, we use the common multi-label cross-entropy loss as the classification loss ($\mathcal{L}_{cls}$). More importantly, we design three losses to optimize the class activation map $\mathcal{M}_{fg}$ through double decoupling learning ($\mathcal{L}_{d2}$) and refinement constraints ($\mathcal{L}_{ma}$ and $\mathcal{L}_{oe}$).

1) Double Decoupled Learning. First, we achieve category decoupling by optimizing each channel (each channel corresponds to a specific category) of the class activation map individually, *i.e.* each $\mathcal{M}^c_{fg}$ and the corresponding $\hat{Y}^c_{fg}$ contains only one category of objects. Then, we adopt a double decoupling loss $\mathcal{L}_{d2}$ to achieve category decoupling and foreground-background decoupling learning for class activation map. Specifically, we construct the one-hot encoding Y^c (only the c category is assigned to 1, and others are assigned to 0) as the ground truth of the c-th decoupling category. Inspired by the triplet loss [10] in metric learning, we design $\mathcal{L}_{d2}$ to make $\hat{Y}^c_{fg}$ and Y^c closer, while make $\hat{Y}^c_{fg}$ and Y^c more distant in metric space. To accommodate the classification task, we use the softmax cross-entropy loss ($\mathcal{L}_{sce}$) but not the cosine distance adopted by the triplet loss [10] as a metric between classification prediction vectors.

Hence, $\mathcal{L}_{d2}$ is defined as:

$$\mathcal{L}_{d2} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\texttt{ReLu}(1-\beta\mathcal{L}_{sce}(\hat{Y}^c_{bg}, Y^c)+\mathcal{L}_{sce}(\hat{Y}^c_{fg}, Y^c)), \tag{5}$$

where β is a weighting factor to adjust the effect of $\mathcal{L}_{sce}$, which is empirically set to 0.1. The ReLu function guarantee that the loss is positive. $\mathcal{L}_{d2}$ encourages $\mathcal{M}_{fg}$ to include foreground object regions, while $\mathcal{M}_{bg}$ to exclude foreground object regions, making the class activation map have a good foreground-background division.

2) Refinement constraints. We also design two simple yet effective refinement constraints to further improve the quality of the class activation map $\mathcal{M}_{fg}$.

First, we find that if $\mathcal{M}_{fg}$ contains a too large area, $\mathcal{L}_{d2}$ will also drop low, so it is necessary to impose a constraint on the area of the $\mathcal{M}_{fg}$ in order to avoid such a locally optimal solution. To obtain a more compact foreground region, a common choice is to directly minimize the area of $\mathcal{M}_{fg}$ (*i.e.* $\mathcal{L}_{area}$ in Eq. 7), as in [30, 41]. However, we find that such an area constraint is in contradiction with $\mathcal{L}_{sce}(\hat{Y}^c_{fg}, Y^c)$ term of $\mathcal{L}_{d2}$ obviously and can cause a tendency to be suboptimal during optimization, *i.e.* one loss is over-optimized, but the other loss is under-optimized. To alleviate this problem, we propose a flexible margin-wise area loss $\mathcal{L}_{ma}$ to stably optimize the area of the foreground region without affecting the optimization of other losses. Specifically, the margin-wise area loss $\mathcal{L}_{ma}$ can be defined as:

$$\mathcal{L}_{ma} = \texttt{ReLu}(\mathcal{L}_{area} - m) + \lambda\mathcal{L}_{area} \tag{6}$$

where

$$\mathcal{L}_{area} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\sum_{i}^{H}\sum_{j}^{W}\mathcal{M}^c_{fg}(i,j). \tag{7}$$

Here m is the margin threshold, and the foreground area larger than m is fully optimized, empirically set to 0.2. The λ is the weight of the area constraint when the foreground area ratio is smaller than m in order to provide a continuous area constraint throughout the optimization process, which is empirically set to 0.05. Specifically, the loss weight of $\mathcal{L}_{area}$ is $(1+\lambda)$ when the sample foreground area ratio for more than m and the weight is λ when the area ratio less than m. During the training, $\mathcal{L}_{d2}$ can be well optimized together with $\mathcal{L}_{ma}$ and constrained to each other but without interrupted.

Second, we propose an overlap elimination loss ($\mathcal{L}_{oe}$) to alleviate the conflicts that exist between different categories of class activation maps. This conflict is caused by the non-exclusive activation [7] of the multi-label classifier and leads to the misclassification of foreground objects pixels. Specifically, $\mathcal{L}_{oe}$ aims to reduce the overlapping regions between different class activation maps shown in Figure 4, which is defined as:

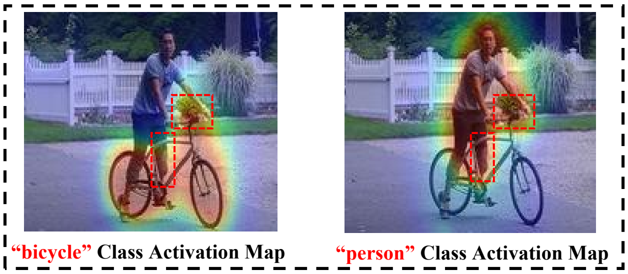


Figure 4. Schematic diagram of $\mathcal{L}_{oe}$. The red boxes indicate the overlapping region of the "bicycle" and "person" class activation maps.

$$\mathcal{L}_{oe} = \sum_{i}^{H}\sum_{j}^{W}\texttt{ReLu}\Big(\big(\sum_{c\in\mathcal{C}}\mathcal{M}^c_{fg}(i,j)\big) - 1\Big). \tag{8}$$

Notice that after summing the class activation maps, only the overlapping regions will have activation values greater than 1, and we reduce the overlapping regions by penalizing the regions greater than 1.

3.2.2 Generate Pseudo-labels

In this section, we describe how to generate pseudo-labels from the class activation map derived from D2CAM. First, we transform the class activation map $\mathcal{M}_{fg}$ obtained by Section 3.2.1 into a *trimap* $T \in \mathbb{R}^{H\times W}$ by setting two thresholds ε_{fg} and ε_{bg}. Specifically, *trimap* T is generated as follows:

$$T(i,j) = \begin{cases} 1\ (foreground) & \mathcal{M}_{fg}(i,j) \geq \varepsilon_{fg} \\ 0\ (background) & \mathcal{M}_{fg}(i,j) \leq \varepsilon_{bg} \\ 0.5\ (unknown) & \varepsilon_{bg}(i,j) < \mathcal{M}_{fg} < \varepsilon_{fg} \end{cases}, \tag{9}$$

where $i \in \mathbb{R}^H$ and $j \in \mathbb{R}^W$. Then, we input the obtained *trimap* T and image I into the image matting algorithm KNN matting [5] for solving alpha matte α, and use alpha matte α as the pseudo-label. In addition, as a common option, our method can cascade some refinement methods (*e.g.* DenseCRF [19] and IRN [1]) to obtain higher quality pseudo-labels.

In fact, our Mat-Label pipeline has only two additional hyperparameters, *i.e.* ε_{fg} and ε_{bg}, than the pure CAM-based solution [47]. The clear distinction between foreground and background regions in our D2CAM (Figure 2 (c)) leads to a robust and insensitive selection of ε_{fg} and ε_{bg}. We empirically set ε_{fg} and ε_{bg} to 0.75 and 0.2, respectively.

4. Experiments

4.1. Experimental Setup

Experiments are conducted on two publicly available and popular datasets, PASCAL VOC 2012 [11] and MS COCO 2014 [32]. The PASCAL VOC 2012 dataset includes 20 foreground and background categories. It has three subsets: training, validation, and test, with 1464, 1449, and 1456 images, respectively. Following previous works [23, 25, 45, 7, 30], we also used the augmented train set with 10,582 training images provided by [12]. The MS COCO 2014 dataset has 80 foreground categories and one background category, including approximately 82K training images and 4K validation images. Note that during the training phase, we use only the image-level category labels of both datasets for supervision, which is the most challenging setting in WSSS. During the testing phase, we evaluate our method on 1,449 validation images and 1,456 test images from the PASCAL VOC dataset, as well as on 40,504 validation images from the MS COCO dataset with ground truth segmentation masks. According to convention, the mean intersection-over-union (mIoU) [33] is used as the evaluation metric.

4.2. Implementation Details

Training. Following [1, 25, 45, 30], we use ResNet-50 [13] as the classification network backbone ($\mathcal{N}_1$ and $\mathcal{N}_2$ in Section 3.2.1). The total loss $\mathcal{L}_{total}$ definition is $\mathcal{L}_{total} = \mathcal{L}_{cls} + \gamma_1 \mathcal{L}_{d2} + \gamma_2 \mathcal{L}_{ma} + \gamma_3 \mathcal{L}_{oe}$, where γ_1, γ_2 and γ_3 are $0.5, 2$ and 0.5 , respectively. Inspired by [41], we truncate the gradients of $\mathcal{N}_1$ and $\mathcal{N}_2$ in the computation of $\mathcal{L}_{d2}$ to allow the network to focus on optimizing the generator $\varphi(\cdot)$. Our D2CAM is implemented based on PyTorch and runs on a PC with a single Nvidia RTX A6000 GPU with the batch size 16. We deploy the SGD optimizer with an initial learning rate $6e^{-3}$ and decays according to the polynomial schedule. We train 8 and 12 epochs on PASCAL VOC and MS COCO datasets, respectively.

Evaluation. Following [1, 25, 7, 30], multi-scale class activation maps $\{0.5,\ 1.0,\ 1.5,\ 2.0\}$ are averaged and then output. The common refinement algorithms DenseCRF [19] and IRN [1] are used as the initial label post-processing and run with default parameters. As in the previous works [7, 30], we retrain the IRN according to the original settings using the initialized pseudo-labels we generated. For the final semantic segmentation step, we use the PyTorch implementation of DeepLab-v2-ResNet101[1].

4.3. Quality of Pseudo-labels

Table 1 reports the mIoU scores of our Mat-Label and its variants compared with recent WSSS pseudo-labels generation methods on the PASCAL VOC 2012 train set [11]. Here, the Seed means initialized pseudo-labels, *w/* CRF means using DenseCRF [19] refinement, and *w/* IRN means using IRN [1] refinement. **Ablation Study:** D2CAM *only* means that the pseudo-labels are generated directly using the class activation map of our D2CAM, while Mat-Label *w/* ReCAM [7] means that the *trimap* is generated using the recent ReCAM [7] to be applied to our Mat-Label pipeline. Note that we have achieved significantly better performance than the most recent ESOL [30] using only the D2CAM (58.0% *vs* 53.6%). With the help of our Mat-Label pipeline, ReCAM [7] has been significantly improved compared to the original version (+1.5%, +2.7%, +1.3%). Our Mat-Label obtained best results (62.3%, 65.8%, 72.9%) after applying the D2CAM specially designed for it, demonstrating that our method is the current optimal choice for weakly supervised *trimap* generation. **Comparison Study:** Both in the quality comparison of the initial seed and the refined pseudo-label, our Mat-Label surpasses the current state-of-the-art methods with a large margin. Specifically, our optimal results for Mat-Label exceed the recent ESOL [30] and ReCAM [7] by 4.2% and 3.2%, respectively.

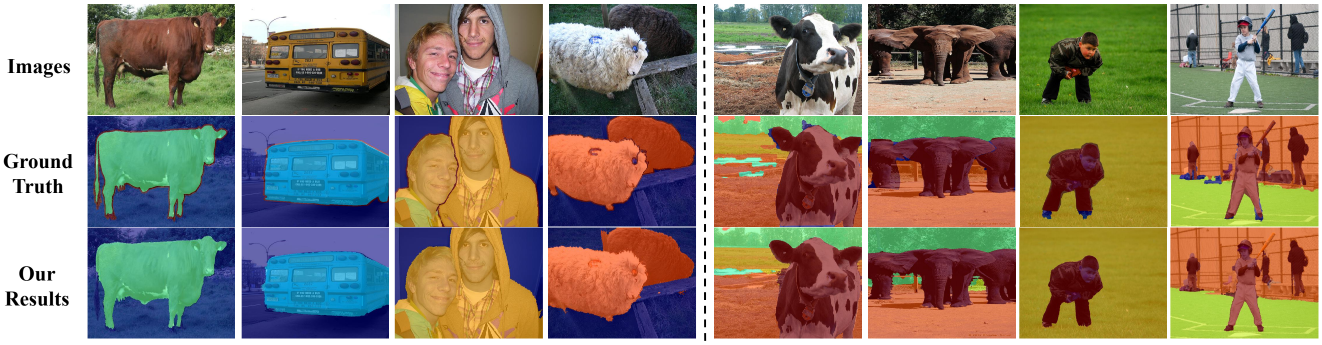
Table 1. Comparison of pseudo-labels quality (mIoU %) on PASCAL VOC 2012 train set [11]. Here * denotes the results of our own implementation. The best results marked with **bold**.

Methods	PASCAL VOC		
	Seed	*w/* CRF [19]	*w/* IRN [1]
CAM $_{CVPR'16}$ [47]	48.0	–	–
IRN $_{CVPR'19}$ [1]	48.8	53.7	66.3
SC-CAM $_{CVPR'20}$ [2]	50.9	–	–
BES $_{ECCV'20}$ [4]	50.4	–	67.2
CONTA $_{NeurIPS'20}$ [45]	48.8	–	67.9
CDA $_{ICCV'21}$ [38]	50.8	58.4	67.7
RIB $_{NeurIPS'21}$ [22]	56.5	62.9	–
L2G $_{CVPR'22}$ [17]	56.2	–	–
ReCAM $_{CVPR'22}$ [7]	54.8	60.4*	69.7*
ESOL $_{NeurIPS'22}$ [30]	53.6	61.4	68.7
D2CAM *only*	58.0	63.9	71.4
Mat-Label *w/* ReCAM [7]	56.3	63.1	71.0
Mat-Label (ours)	**62.3**	**65.8**	**72.9**

4.4. Weakly Supervised Semantic Segmentation

Table 2 and Table 3 report the segmentation performance of the advanced WSSS methods on the PASCAL VOC 2012 [11] and MS COCO 2014 [32] datasets. Some of these methods improve the pseudo-labels generation while others refine the pseudo-labels, but they all use the same segmentation network setup (DeepLab-V2 [3]) to evaluate performance. Table 2 shows the mIoU score on PASCAL VOC validation and test set under image category label (*I.*) or image category label & saliency maps (*I.*+ *S.*) supervision. Following ESOL [30], we combine the saliency map

[1] https://github.com/kazuto1011/deeplab-pytorch



(a) PASCAL VOC 2012 dataset **(b) MS COCO 2014 dataset**

Figure 5. Some qualitative results on the PASCAL VOC 2012 and MS COCO 2014 datasets. We can see that our method achieves accurate final semantic segmentation results under image-level label supervision only.

Table 2. Comparison of the mIoU (%) with state-of-the-art approaches on PASCAL VOC 2012 validation and test sets. All the segmentation results are based on the DeepLab-V2 [3].

Mthods	Sup.	Val (%)	Test (%)
SEAM $_{CVPR'20}$ [39]	*I.*	64.5	65.7
SC-CAM $_{CVPR'20}$ [2]	*I.*	66.1	65.9
BES $_{ECCV'20}$ [4]	*I.*	65.7	66.6
CONTA $_{ECCV'20}$ [45]	*I.*	66.1	66.7
CDA $_{ICCV'21}$ [38]	*I.*	66.1	66.8
OC-CSE $_{ICCV'21}$ [21]	*I.*	68.4	68.2
RIB $_{NeurIPS'21}$ [22]	*I.*	68.3	68.6
AdvCAM $_{TPAMI'22}$ [24]	*I.*	68.1	68.0
VWL $_{IJCV'22}$ [37]	*I.*	69.2	69.2
W-OoD $_{CVPR'22}$ [26]	*I.*	70.7	70.1
SIPE $_{CVPR'22}$ [6]	*I.*	68.8	69.7
RECAM $_{CVPR'22}$ [7]	*I.*	68.5	68.4
ESOL $_{NeurIPS'22}$ [30]	*I.*	69.9	69.3
D2CAM *only*	*I.*	71.2	70.7
Mat-Label *w/* ReCAM [7]	*I.*	70.3	70.0
Mat-Label (ours)	*I.*	**73.0**	**72.7**
DSRG $_{CVPR'18}$ [15]	*I.+S.*	61.4	63.2
FickleNet $_{CVPR'19}$ [23]	*I.+S.*	64.9	65.3
NSROM $_{CVPR'21}$ [43]	*I.+S.*	70.4	70.2
EPS $_{CVPR'21}$ [28]	*I.+S.*	70.9	70.8
AdvCAM $_{TPAMI'22}$ [24]	*I.+S.*	71.3	71.2
ReCAM $_{CVPR'22}$ [7]	*I.+S.*	71.8	72.2
L2G $_{CVPR'22}$ [17]	*I.+S.*	72.1	71.7
RCA $_{CVPR'22}$ [48]	*I.+S.*	72.2	72.8
ESOL $_{NeurIPS'22}$ [30]	*I.+S.*	71.1	70.4
Mat-Label (ours)	*I.+S.*	**73.3**	**74.0**

generated by NSROM [43] with our pseudo ground-truth masks to supervise the segmentation network (*I.+ S.*). **Ablation Study:** On both the PASCAL VOC 2012 and MS COCO 2014 datasets, it can be seen that both of our non-optimal variants achieve competitive segmentation performance, and the performance lead is further extended when the full Mat-Label is used. **Comparison Study:** In Table 2, we can see that our method outperforms the most recent ESOL [30] by 3.1% and 3.4% on the validation set and test set, respectively. Our method achieved 73.3% and 74.0% mIoU scores in the PASCAL VOC 2012 validation and test sets after imposing saliency map supervision, consistently outperforming all other saliency map based methods. Table 3 illustrates the segmentation performance on the MS COCO dataset compared with other state-of-the-art approaches. Our method achieves the best result with 2.7% improvement compared to the second-best method ReCAM [7]. In addition, Figure 5 presents some examples of the final semantic segmentation results on both PASCAL VOC 2012 and MS COCO 2014 datasets, demonstrating that our results are close to the ground truth segmentation masks.

Table 3. Comparison of the mIoU (%) with state-of-the-art approaches on MS COCO validation set

Methods	Sup.	Val (%)
IRN $_{CVPR'19}$ [1]	*I.*	41.4
ADL $_{TPAMI'20}$ [8]	*I.*	30.8
SEAM $_{CVPR'20}$ [39]	*I.*	32.8
CONTA $_{NeurIPS'20}$ [45]	*I.*	33.4
OC-CSE $_{ICCV'21}$ [21]	*I.*	36.4
SIPE $_{CVPR'22}$ [6]	*I.*	40.6
ReCAM $_{CVPR'22}$ [7]	*I.*	42.9
ESOL $_{NeurIPS'22}$ [30]	*I.*	42.6
D2CAM *only*	*I.*	44.0
Mat-Label *w/* ReCAM [7]	*I.*	43.8
Mat-Label (ours)	*I.*	**45.6**

4.5. Ablation Study on D2CAM

Table 4 shows the ablation study results of our D2CAM on the PASCAL VOC 2012 train set. It can be observed that higher quality class activation map (performance +5.7% over the pure CAM [47]) is obtained when we impose $\mathcal{L}_{d2}$ loss to introduce double decoupled learning. Note that after applying $\mathcal{L}_{area}$ (Eq. 7), there is a performance degradation due to falling into local optima. In contrast, our D2CAM achieves the higher mIoU score after replacing $\mathcal{L}_{area}$ with our proposed $\mathcal{L}_{ma}$ (Eq. 6). Finally, the quality of the class activation map is further improved by applying $\mathcal{L}_{oe}$ later, due to the imposition of constraints on the overlapping regions between different categories.

Table 4. D2CAM ablation study on PASCAL VOC 2012 train set

Methods	mIoU (%)
CAM [47]	48.0
$\mathcal{L}_{d2}$	53.7
$\mathcal{L}_{d2}$ + $\mathcal{L}_{area}$	50.4
$\mathcal{L}_{d2}$ + $\mathcal{L}_{ma}$	56.8
$\mathcal{L}_{d2}$ + $\mathcal{L}_{ma}$ + $\mathcal{L}_{oe}$	**58.0**

5. Conclusions

We propose an interesting and promising pipeline called Mat-Label to treat WSSS pseudo-labels generation as the image matting task. To achieve this goal, we propose a double decoupled class activation map (D2CAM) for generating *trimap* with good foreground-background division under weak supervision. Extensive experiments validate that our Mat-Label significantly improves the quality of the pseudo-label, exhibiting state-of-the-art performance both on the PASCAL VOC 2012 and MS COCO 2014 datasets. Furthermore, our D2CAM can independently outperform the existing CAM-based WSSS methods. We hope that our Mat-Label solution can shed new light on WSSS pseudo-labels generation and other weakly supervised tasks.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grants $U21A20515$, 62271074, $U2003109$, 62171321, 62071157, 62162044,61972459, and 62102414), the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Sciences (Grant No. LSU-KFJJ-2021-05), and the Open Projects Program of National Laboratory of Pattern Recognition.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019.

[2] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020.

[3] LC Chen, G Papandreou, I Kokkinos, K Murphy, and AL Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[4] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020.

[5] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[6] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4288–4298, June 2022.

[7] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022.

[8] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4256–4271, 2020.

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[10] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7014–7023, 2018.

[16] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2070–2079, 2019.

[17] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16886–16896, 2022.

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 695–711. Springer, 2016.

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

[20] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533, 2017.

[21] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6974–6983. IEEE, 2021.

[22] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021.

[23] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019.

[24] Jungbeom Lee, Eunji Kim, Jisoo Mok, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[25] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021.

[26] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022.

[27] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1854–1862, 2021.

[28] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021.

[29] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007.

[30] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2209.07761*, 2022.

[31] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[34] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3265–3274. IEEE, 2019.

[35] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021.

[36] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13673–13682. Computer Vision Foundation / IEEE, 2020.

[37] Lixiang Ru, Bo Du, Yibing Zhan, and Chen Wu. Weakly-supervised semantic segmentation with visual words learning and hybrid pooling. *International Journal of Computer Vision*, 130(4):1127–1144, 2022.

[38] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised se-

mantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021.

[39] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[40] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7268–7277, 2018.

[41] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14228–14237. IEEE, 2022.

[42] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C^2 AM: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 979–988. IEEE, 2022.

[43] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2623–2632, 2021.

[44] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 326–344. Springer, 2022.

[45] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[46] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018.

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

[48] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4299–4309, June 2022.

[49] Lei Zhu, Qi She, Qian Chen, Xiangxi Meng, Mufeng Geng, Lujia Jin, Zhe Jiang, Bin Qiu, Yunfei You, Yibao Zhang, et al. Background-aware classification activation map for weakly supervised object localization. *arXiv preprint arXiv:2112.14379*, 2021.