IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 20, 2020, accepted October 30, 2020, date of publication November 3, 2020, date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035707

Weakly-Supervised Semantic Segmentation With Regional Location Cutting and Dynamic Credible Regions Correction

MINGSI TONG[1], (Member, IEEE), WENCONG LI[1], XINYANG REN[1], (Member, IEEE), XINGHU YU[2,3], AND WEIYANG LIN[1]

[1]Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin 150001, China
[2]Ningbo Institute of Intelligent Equipment Technology, Ningbo 315000, China
[3]Harbin Institute of Technology (HIT), Harbin 150090, China

Corresponding author: Weiyang Lin (wylin@hit.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61973099 and Grant 61933008, and in part by the National Postdoctoral Program for Innovative Talents under Grant BX201600044.

ABSTRACT Weakly-supervised semantic segmentation is a challenging task as it outputs pixel-level predictions from weaker labels. Segmentation with weaker labels is an important research area since it can significantly reduce human annotation efforts by associating high-level semantic to low-level appearance. In this article, we propose a novel *Regional location Cutting and Dynamic credible regions Correction* (RCDC) approach for weakly-supervised semantic segmentation. Only image-level labels are needed and it can take less time for manual annotation. Starting with the weak localization of classification network, our cutting approach combines the weak coverage with the traditional cutting method to obtain the pseudo-labels of around 50% ground truth. Then, our dynamic credible regions correction approach adjusts the loss function during the training to preserve the regions that have the superior performance of each iteration. It can further enhance the pseudo-labels for better segmentation results. Finally, with the fully-connected CRF and the retraining method, our approach obtains a competitive performance on the PASCAL VOC 2012 dataset.

INDEX TERMS Semantic segmentation, weak supervision, regional location cutting, dynamic credible regions correction.

I. INTRODUCTION

With the development of *Convolutional Neural Networks* (CNNs) recently, computer vision research has made immense progress to improve our lives. A great number of computer vision tasks now can be solved by training the CNNs with datasets that have labels. And a large amount of fully fully-annotated is the key to training results. Generally, there are three main areas in computer vision: classification with image-level labels [1]–[5], detection with bounding box labels [6]–[8], segmentation with pixel-level labels [9]–[13]. For Image-level labels, every training image has its image category. Bounding box labels show every bounding box of every object in the image and the pixel-level labels mark every pixel's category of the images. Manual annotation now is the only way to make a large amount of fully-annotated images and it always time-consuming work. Generally, it will take 1 second for an image-level label, the bounding box label will take about 7 seconds per image on average and it may take more than one and a half hours to make a pixel-level labeled image [14]. It is obvious that labeling the segmentation training set takes dozens or even hundreds of times than image-level and bounding box labels. So for some computer vision tasks, especially segmentation tasks, it is a great challenge to make the fully-annotated images for training.

Compared with image-level and bounding box labels, the pixel-level labels needs too much labor cost. If we can realize semantic segmentation just with weaker labels such as image-level and bounding box labels, a lot of manpower and timing costs will be saved. This is of great significance to the engineering application of semantic segmentation. Therefore, a great number of researchers are now focusing on training the segmentation network with unlabelled or

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Cheng.

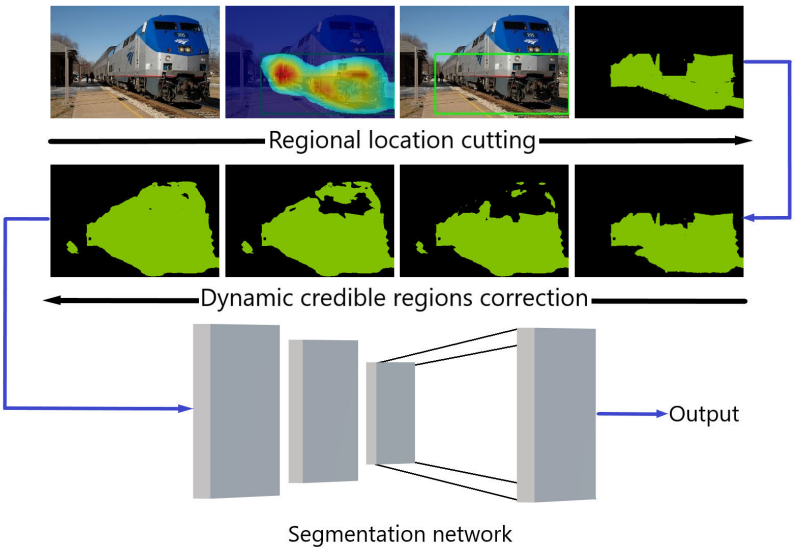

FIGURE 1. RCDC: The top row shows the pseudo-labels from the image cutting operation. The middle row shows with the dynamic credible regions training correction operation, the pseudo-labels are improved. The bottom row shows the final predicts output from the segmentation network.

weakly labeled images [15]–[18]. It can reduce the workload of making annotated images dramatically and increase efficiency.

Weakly-supervised image segmentation is now a promising direction in computer vision research by just using image-level or bounding box labeled datasets to train the segmentation network. Compared with pixel-level labels, image-level labels and bounding box labels have less supervision information, so-called weakly-supervised image segmentation. *Classification Activation Maps* (CAMs) [19] is a very important research work for weakly-supervised image segmentation. Oquab [19] found that an image classification network can show the location of the object in the image, and it provides the possibility for training segmentation network with image-level labels. Bounding box labels provide more information than image-level labels including the location of the objects and the bounding box of each object. There have been a great number of works that focus on weakly-supervised image segmentation. SEC [20] proposed by Alexander *et al.* uses CAM to obtain the seeds of the objects and then expands it to the ground truth. It has been proved to be an effective method for image-level semantic segmentation. But the seeds obtained from CAM may just contain less than 30% right pixels of ground truth which limits the final segmentation result. SDI [21] introduced by Anna Khoreva shows that retraining the segmentation net is an active method to improve the weak labels. But with some bad labels getting better during the retraining, some good labels will become worse.

In this article, after investigating a great number of previous researches of weakly-supervised semantic segmentation, we put forward two main principles that almost all of those methods are based on:

- **Getting better pseudo-labels from weakly labeled images**: We know that the input of the segmentation network is the pixel-level labeled images so the input of weakly-supervised segmentation is called pseudo-labels. It is obvious that the more accurate the pseudo-label is, the better the segmentation result will be. So getting better pseudo-labels is an important aspect of weakly-supervised segmentation.
- **Improving loss function, training method, or network structure for better segmentation result**: The pseudo-label sometimes is just a little part of the ground truth and it is a significant task to improve the segmentation result during training. In some researches, some new loss functions are put forward for better segmentation and some new network structures have been proved are effective for weakly-supervised segmentation. There are also some new training methods for weakly-supervised segmentation.

Generally, by following these two principles, a nice weakly-supervised segmentation result can be reached. In this article, our research of weakly-supervised semantic segmentation also focuses on these two main principles. By following the two principles we proposed above, we propose a weakly supervised semantic image segmentation method with *Regional location Cutting and Dynamic credible regions Correction* (RCDC). It is proved to be an effective method for both image-level labels and bounding box labels. The Fig. 1 shows the architecture of RCDC.

The better pseudo-labels are the better the segmentation result will be. Considering the pseudo-labels that made by CAM are about 30% of the ground truth, we expand the weak localization and cut with the traditional segmentation method, which can make better pseudo-labels with

image-level images than just CAM. Inspired by WILDCAT [22], the heatmap that comes from the classification network can show more complete coverage of the object. Some traditional segmentation algorithms can reach a nice segmentation result with the region that contains the object [23], [24]. Therefore, we can get better pseudo-labels by expanding the coverage of the object and cutting with the traditional segmentation algorithm.

By training the segmentation network with pseudo-labels, the bad pseudo-labels can cause the network overfitting to worse performance. To address this issue, in this article we propose a training dynamic correction loss that can adjust the loss with the confidence degree of the score maps to save the regions of good performance during the training. It contains expanding the pseudo-labels to the ground truth, cutting off redundant labels, and adding the missing labels. We also add a fuzzy element to the loss, which can train the network with the mask of the clear category regions and the unclear category regions. The different losses of certain regions and the uncertain regions can help the network recognize the right regions and abandon the wrong regions.

With our weakly-supervised image segmentation, we can reach the segmentation results just using image-level labels or bounding box labels. It can save significant time for manual annotation. In some engineering applications, our method can effectively reduce costs.

The main contributions of our work are two-fold:

- We propose a regional location cutting method by expanding the location of WILDCAT to cover the whole objects. And with the traditional segmentation method Grabcut, the pseudo-labels we obtain can have almost 50% or 70% right pixels of the ground truth.
- The dynamic credible regions training correction method based on the dynamic region loss function is developed to correct the wrong regions of the pseudo-labels. It can improve the results of weakly-supervised segmentation.

II. RELATED WORK

Fully-supervised and weakly-supervised semantic segmentation networks that are related to our work are introduced in this section.

A. FULLY-SUPERVISED SEMANTIC SEGMENTATION

Fully-supervised segmentation network is a pixel-based *end to end* fully convolutional network which has the input of entire images and predicts pixel-wise outputs. It is a method that needs a great number of pixel-wise annotations. *Fully Convolutional Networks* (FCNs) [25] with skip architecture to produce accurate semantic segmentation that proposed by Long *et al.* lays a foundation for semantic segmentation of convolutional neural network. Semantic segmentation network has made great progress in recent years. *Deeplab network* [26]–[28], [30] designed by Chen *et al.* proposed Encoder-Decoder architecture and Atrous Convolution. It can enlarge the receptive field with a lower stride to produce denser segmentation. Now, most of the segmentation networks are developed from FCNs and Deeplab.

Most segmentation networks such as FCNs and Deeplab are relied on the cross-entropy loss function ℓ as shown in following

$$\ell = -\frac{1}{p}\sum_{i=1}^{p}\log\left(f_i(c)\right) \tag{1}$$

where p is the number of pixels in the image or minibatch. $\mathcal{C}$ is the set of all the categories. $c \in \mathcal{C}$ is the ground truth class of the pixel. $f_i(c)$ is predict probability of the ground truth of pixel i. Generally, $f_i(c)$ is got from the unnormalized scores $F_i(c)$ that output from the network through a softmax unit.

$$f_i(c) = \frac{e^{F_i(c)}}{\sum\limits_{c' \in \mathcal{C}} e^{F_i(c')}} \quad \forall i \in [1, p], \forall c \in \mathcal{C} \tag{2}$$

B. WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

A great number of pixel-wise annotations which are very time-consuming are required for the fully-supervised semantic segmentation. So the weakly-supervised semantic segmentation with weak labels came out recently including bounding box labels and image-level labels.

Because image-level weakly-supervised semantic segmentation only has the classification labels, most of the segmentation methods are based on CAM which can identify a region of the ground truth using the heatmap of the classification network. And the regions of CAM are always used as initial objects segmentation regions. Some extra methods are used to improve the segmentation results. Kolesnikov *et al.* (2016) proposed a new loss of seed, expand and constrain [20] which can seed the initial region with CAM and expand it and constrain it to the ground truth. Wei *et al.* (2017) proposed a training method of region mining with adversarial erasing. It uses CAM to locate the object and erase it, then locates it again to locate different object features [29]. Wang *et al.* (2018) proposed a new architecture which contains a region net a pixel net to improve the segmentation result [31]. SSNet(2019) by Zeng *et al.* [33] is a weakly-supervised semantic segmentation method with the supervision of image-level labels and Saliency labels. DSRG(2018) by Huang *et al.* [32] and GAIN(2018) by Li *et al.* [35] are also efficient image-level weakly-supervised semantic segmentation methods.

The bounding box labels provide the category and the bounding box of every object which can locate every object just from labels. And most segmentation methods with bounding box labels are combined with some traditional segmentation methods. By using the traditional segmentation to every bounding box, the pseudo-labels of about 70 percent right pixels can be made for future segmentation. Anna Khoreva(2017) proposed a method called Simple Does It [21] which is a new approach that does not require modification of the segmentation training procedure and uses bounding boxes to design the input labels. Most of the segmentation methods

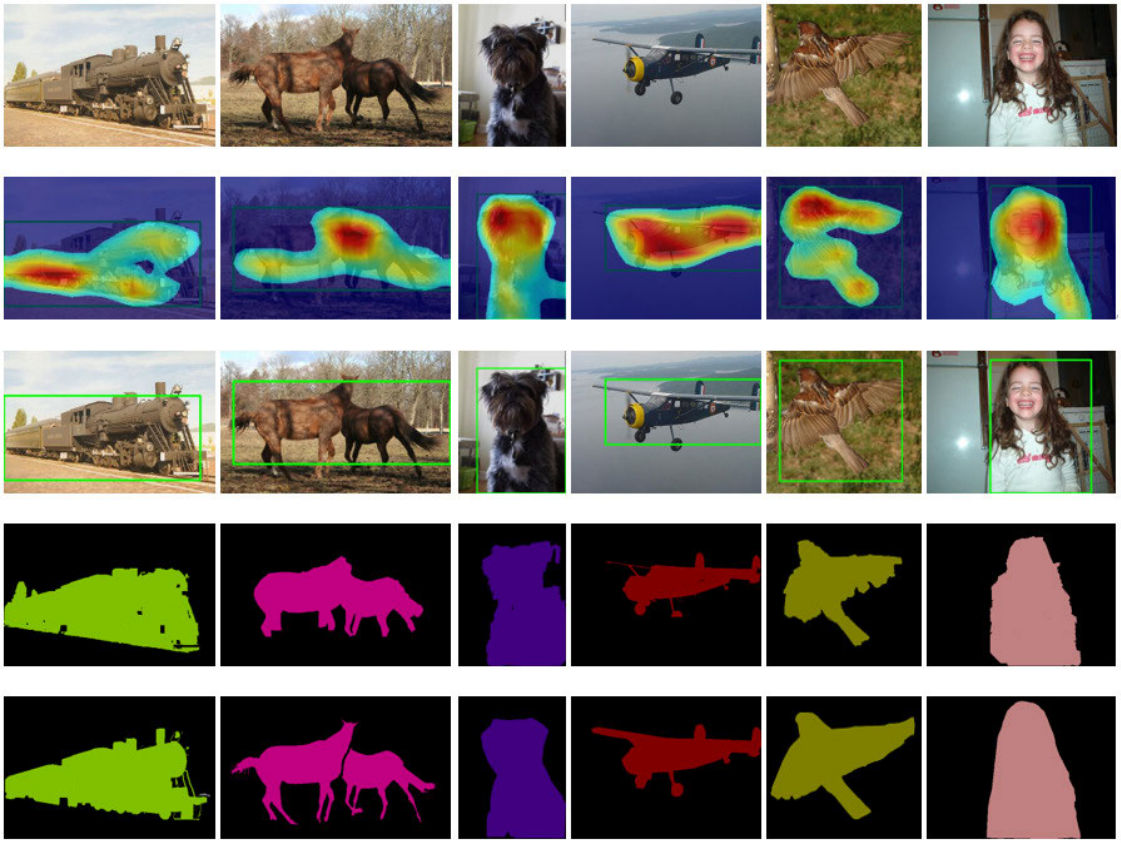

FIGURE 2. Regional location cutting: The first row shows the original images. The second row shows the location map $L_i(x, y)$. The third row shows the weak bounding boxes from image-level labels. The fourth row shows the pseudo-labels made from the weak bounding boxes. The last row shows the ground truth of segmentation.

of bounding box labels are focused on making better pseudo-labels, while in this article, we focus on not only making better pseudo-labels but also improving the segmentation result during training.

III. METHODS

A. REGIONAL LOCATION CUTTING

The bounding box labels provide both the categories and the bounding boxes of the objects which can locate every object and provide the coverage of every object just from the label, while the image-level labels just provide the categories of the objects. By training the classification network with image-level labels, the heatmap which can locate the position of the objects will be obtained from the output of the network. A weak coverage of the object can be obtained by processing the heatmap. With the weak coverages of the image-level labels and the coverages of the bounding box labels, we can obtain the pseudo-labels for future segmentation.

1) COVERAGES FROM IMAGE-LEVEL LABELS AND BOUNDING BOX LABELS

The bounding box labels provide the bounding box of each object where the whole object is surrounded by the bounding box and we can get the coverages from the labels. But for image-level labels, we only know the categories of the images so it is a tough work to yield the weak coverages from the image-level images. Previous studies have shown that by combing the score maps with the weights in the network of different objects, the output of the classification network can locate the objects despite having no objects' location annotation at training. CAM is a useful method that is used for making the initial pseudo-labels in most weakly supervised semantic image segmentation procedures. But it can only locate some significant features of the objects which sometimes are just a little part of the objects, while there is a large part of the objects the pseudo-labels can not cover. The WILDCAT is a localization method that not only focuses on the significant features but also the other features of the objects, so it can cover more of the objects compared with the CAM method.

Consider (x, y) is the coordinate of the pixel in the image. $u_j(x, y)$ and w_{ij} denote the feature maps of the last layer and the weight of each feature map respectively, where i stands for the i^{th} class of n classes and j is the j^{th} feature map of r feature maps. The score map $s_i(x, y)$ of class i from WILDCAT is the weighted sum of the feature maps.

$$s_i(x, y) = \sum_{j=1}^{r} w_{ij} \cdot u_j(x, y) \tag{3}$$

Resize the score map to the size of original image, we can get the score map $\hat{s}_i(x, y)$ which shows the location of objects, and the objects of class i can be located by the location map $q_i(x, y)$ defined as following

$$q_i(x,y) = \begin{cases} \hat{s}_i(x,y), & \hat{s}_i(x,y) > \text{tr and label}(i) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where tr is the threshold we choose to limit the size of the coverages. label(i) is the muti-classification label function, label(i) $= 1$ means the image has the object of class i, while label(i) $= 0$ means the image doesn't has the object of class i. We consider all the regions that $q_i(x, y) \neq 0$ are the coverages of the objects. Because the shape of the coverages are irregular we reshape the irregular regions to the weak boxes $\mathcal{I}_{\text{box}} = \{(x, y)|x \in [x_{\min}, x_{\max}], y \in [y_{\max}, y_{\max}]\}$ which are the external rectangles of the regions.

2) PSEUDO-LABELS FROM THE BOXES

After the operation mentioned above, both bounding box labels and image-level labels convert to the coverages of boxes $\mathcal{I}_{\text{box}} = \{(x, y)|x \in [x_{\min}, x_{\max}], y \in [y_{\max}, y_{\max}]\}$ and the whole objects are contained in the boxes. GrabCut [34] is an established traditional method to cut an object from its bounding box. For each annotated box, generating the foreground of the location map $\mathcal{L} = \{(x, y)|q_i(x, y) \neq 0, (x, y) \in \mathcal{I}_{\text{box}}\}$ and the output $\mathcal{G}$ of the GrabCut, we set the pixels of the foreground to the box object class. If $\{\mathcal{L} \cup \mathcal{G}\} \setminus \{\mathcal{L} \cap \mathcal{G}\}$ is too small, there is a big difference between $\mathcal{L} \cap \mathcal{G}$ and $\mathcal{L} \cup \mathcal{G}$, we consider the output of the GrabCut is not the cut of the object and it has a huge difference from the ground truth. In order to ensure the accuracy of segmentation, the pixels in the box are set as background and in the next section, the missing objects will be complemented. The pseudo-labels from the boxes can cover about 50% right pixels of the ground truth which is superior to the CAM. Algorithm 1 describes the regional location cutting and Fig. 2 shows the pseudo-labels results achieved by the proposed regional location cutting procedure.

Algorithm 1 Regional Location Cutting

Require: The number of categories: n; The number of feature maps: r; The feature map of the last layer: $u_j(x, y)$; The weight of each feature map:w_{ij}; The muti-classification label: label(i); The threshold: tr; The set of all the pixels: $\mathcal{I}_{\text{img}}$

Ensure: pseudo-label: pl

for $i = 1$ to n **do**
  $s_i(x, y) = \sum_{j=1}^{r} w_{ij} \cdot u_j(x, y)$
  Calculate the score map of image size $\hat{s}_i(x, y)$ with $s_i(x, y)$
  while $(x, y) \in \mathcal{I}_{\text{img}}$ **do**
    if $\hat{s}_i(x, y) >$ tr and label(i) $= 1$ **then**
      $q_i(x, y) = \hat{s}_i(x, y)$
    else
      $q_i(x, y) = 0$
    end if
  end while
  Calculate the weak box $\mathcal{I}_{\text{box}}$ with $q_i(x, y)$
  Calculate the pseudo-label of class i pl_i with $\mathcal{I}_{\text{box}}$ and GrabCut
end for
Calculate the pseudo-label pl with pl_i
return pl

B. DYNAMIC CREDIBLE REGIONS TRAINING CORRECTION

Some of the pseudo-labels we get from image cutting are not approximate to the ground truth. Some pseudo-labels may have just a part of pixels of the objects and some pixels that are annotated as the objects are not objects of the ground truth. So it is an important work to correct the pseudo-labels to the ground truth. Some works have shown that the segmentation network has some fault tolerance and it can correct the labels to a certain extent during the training. The more right labels are, the better the training correction will be. However, during the training, with the bad pseudo-labels becoming better, some right pseudo-labels may get worse. To avoid this problem and make full use of the correction ability of the segmentation network, we propose a dynamic region loss with two parts: the foreground expand loss and the fuzzy background loss. It can preserve the regions of great performance during the training of the semantic segmentation network.

1) THE DYNAMIC REGION LOSS

Most of the segmentation networks use the cross-entropy loss as the loss function of the networks. In this article, in order to make full use of the great performance of each iteration during training, we propose a dynamic region loss which can change with training.

The image coordinate domain is denoted by $\mathcal{I}_{\text{img}}$ and its pixel coordinate is denoted by $(x, y) \in \mathcal{I}_{\text{img}}$. Assuming n is the total number of the target classes which belong to the foreground of the image, $\mathcal{N} = \{0, 1, 2, \ldots, n\}$ is an identifier set which is correspondingly identified as the background and the target classes, the background is marked by 0 for simplifying the modeling.

For a given $i \in \mathcal{N}$, the i^{th} target class, $h_i(x, y)$ denotes score output from the segmentation network of pixel (x, y) for the i^{th} target class. *Membership function* $m : \mathcal{I}_{\text{img}} \rightarrow \mathcal{N}$ is used to describe which the target includes the given pixel,

$$i = m(x, y) \quad (5)$$

where $i \in \mathcal{N}$ and the pixel $(x, y) \in \mathcal{I}_{\text{img}}$, the Eq. (5) means the pixel (x, y) belongs to the i^{th} target.

$m_{\text{lab}}(x, y)$ and $m_{\text{pred}}(x, y)$ are denoted as the membership function of the pseudo-label and the membership function of the network predict results respectively.

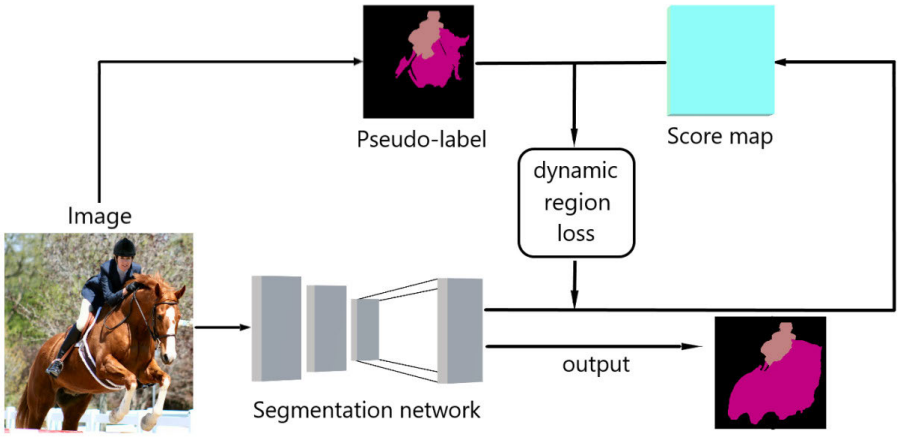

FIGURE 3. The architecture of dynamic credible regions training correction.

$f_i(x, y)$ is the softmax output of $h_i(x, y)$ and the value of $f_i(x, y)$ is between 0 and 1. The value of $f_i(x, y)$ is always considered to be the confidence probability that pixel (x, y) belongs to the i^{th} target class.

$$f_i(x, y) = \frac{\exp(h_i(x, y))}{\sum_{k=0}^{n} \exp(h_k(x, y))} \tag{6}$$

For any given pixel $(x, y) \in \mathcal{I}_{\text{img}}$, let $\alpha = m_{\text{lab}}(x, y)$ and $\beta = m_{\text{pred}}(x, y)$, the dynamic region loss of one pixel (x, y) is

$$p = \begin{cases} 1 - \dfrac{\eta - T}{\eta_{\max} - T}, & \eta - T > 0 \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

$$\ell_{\text{df}}(x, y) = \begin{cases} -\log\left(f_\beta(x, y)\right), & (x, y) \in \mathcal{V}_1 \\ 0, & (x, y) \in \mathcal{V}_2 \\ -\log\left(f_\alpha(x, y)\right), & \text{otherwise} \end{cases} \tag{8}$$

where

$$\begin{aligned} \mathcal{V}_1 = \{(x, y) |\ & \alpha = m_{\text{lab}}(x, y) = 0, \\ & \beta = m_{\text{pred}}(x, y) \neq 0,\ f_\beta(x, y) > p\} \\ \mathcal{V}_2 = \{(x, y) |\ & \alpha = m_{\text{lab}}(x, y) \neq 0, \\ & \beta = m_{\text{pred}}(x, y) = 0,\ f_\beta(x, y) > p\} \end{aligned}$$

η is the iterations and $\eta_{\max}$ is the maximum iteration during training. T is a threshold we set. p is a threshold that changes during the training to select the regions with high confidence.

The pixels in $\mathcal{V}_1$ are the region that has high confidence of the foreground but labeled background in pseudo-label. The first part of $\ell_{\text{df}}(x, y)$ is the foreground expand loss which can expand the foreground with the high confidence regions. It can also add the missing objects to the images.

The pixels in $\mathcal{V}_2$ are the regions that have high confidence of background but labeled foreground in pseudo-label. Because the background is very complex with many kinds of objects that do not belong to the foreground, The second part of $\ell_{\text{df}}(x, y)$ is a fuzzy loss by not giving a clear target for the uncertain background.

Consider N is the number of the pixels in one image. The dynamic region loss of an image L_{df} is

$$L_{\text{df}} = \frac{1}{N} \sum_{(x,y) \in \mathcal{I}_{\text{img}}} \ell_{\text{df}}(x, y) \tag{9}$$

By training the segmentation network with the dynamic region loss, we can improve the pseudo-labels with the performance of each iteration. Algorithm 2 describes the dynamic region loss and Fig. 4 shows the improvement of

Algorithm 2 Dynamic Region Loss

Input: The iteration now and maximum iteration during training:η, $\eta_{\max}$; The threshold: T; The membership function of pseudo-label and network predict: $m_{\text{lab}}(x, y)$, $m_{\text{pred}}(x, y)$; The set of all the pixels: $\mathcal{I}_{\text{img}}$; the output score of class i:$h_i(x, y)$; The number of pixels: N.

Output: dynamic region loss: L_{df}

$L_{\text{df}} = 0$
while $(x, y) \in \mathcal{I}_{\text{img}}$ **do**
  Calculate threshold p with η, $\eta_{\max}$ and T
  Calculate the softmax predict of class i $f_i(x, y)$ with $h_i(x, y)$
  Calculate set $\mathcal{V}_1$, $\mathcal{V}_2$ with $f_i(x, y)$, p, $m_{\text{lab}}(x, y)$ and $m_{\text{pred}}(x, y)$
  if $(x, y) \in \mathcal{V}_1$ **then**
    $\ell_{\text{df}}(x, y) = -\log\left(f_\beta(x, y)\right)$
  else if $(x, y) \in \mathcal{V}_2$ **then**
    $\ell_{\text{df}}(x, y) = 0$
  else
    $\ell_{\text{df}}(x, y) = -\log\left(f_\alpha(x, y)\right)$
  end if
  $L_{\text{df}} = L_{\text{df}} + \ell_{\text{df}}(x, y)$
end while
$L_{\text{df}} = L_{\text{df}}/N$
return L_{df}

FIGURE 4. Dynamic credible regions training correction: The first row shows the images. The second row shows pseudo-labels. The third row shows pseudo-labels after correction. The last row shows the ground truth.

the pseudo-labels with the dynamic credible regions training correction.

IV. EXPERIMENTS

In this section, we first introduce the dataset which is employed, then explain the establishment of a comparative experiment. Finally, we compare our results with other methods presented.

A. DATASET AND SETTINGS

We evaluate the proposed RCDC approach on the PASCAL VOC 2012 segmentation benchmark dataset [36]. PASCAL VOC 2012 segmentation dataset contains three parts: train (1449 images for training), val (1449 images for validation), and test (1456 images for testing). There are 20 object classes and 1 background class in the dataset. The dataset is augmented by the extra annotations provided by [36], resulting in 10582 (trainaug) training images. The performance is measured in terms of pixel *Intersection-Over-Union* (IOU) averaged across the 21 classes.

In the regional location cutting, the backbone network we choose in WILDCAT is Res2net [37] which is a new multi-scale backbone architecture proposed by Shang-Hua Gao (2019). With the results of the experiments, we found Res2net has high performance in weak localization. The weak box of each object in the image consists of the weak localization and bounding box labels which are obtained by regional location cutting. The GrabCut is an effective traditional image cut method, GrabCut is used to make the pseudo-labels from the weak localization boxes in this article. In the dynamic credible regions training correction, the backbone network we choose in segmentation with the dynamic region loss is deeplabV3plus-xception.

Retraining the network can get better segmentation results compared with the result before retraining. It has been proved that fully-connected CRF is an active method to improve the performance of segmentation. We retrain the network three times with fully-connected CRF.

B. RESULTS

In our experiments, for image-level labels, we first train the classification network Res2net with image-level labels. Then the Regional Location Cutting contains weak localization and GrabCut is used to make the pseudo-labels for the segmentation training. We then train the segmentation network *deeplabV3plus* with the dynamic credible regions training correction. Finally, the retraining method is used to improve the final segmentation result. For bounding box level labels, only the method of making pseudo-labels is different from the image-level labels. We just combine the bounding box and the traditional segmentation method GrabCut to make the pseudo-labels.

After training with *PyTorch* implementation of deeplabV3plus, the result of our approach is shown in table 1.

TABLE 1. The result of different results of using different methods by evaluated in PASCAL VOC 2012 val set.

Methods	mIOU
Regional Location Cutting in val set	44.75
baseline(deeplabV3plus + Regional Location Cutting)	56.15
+ Dynamic Region Loss	61.05
+ Retrain and CRF	63.55

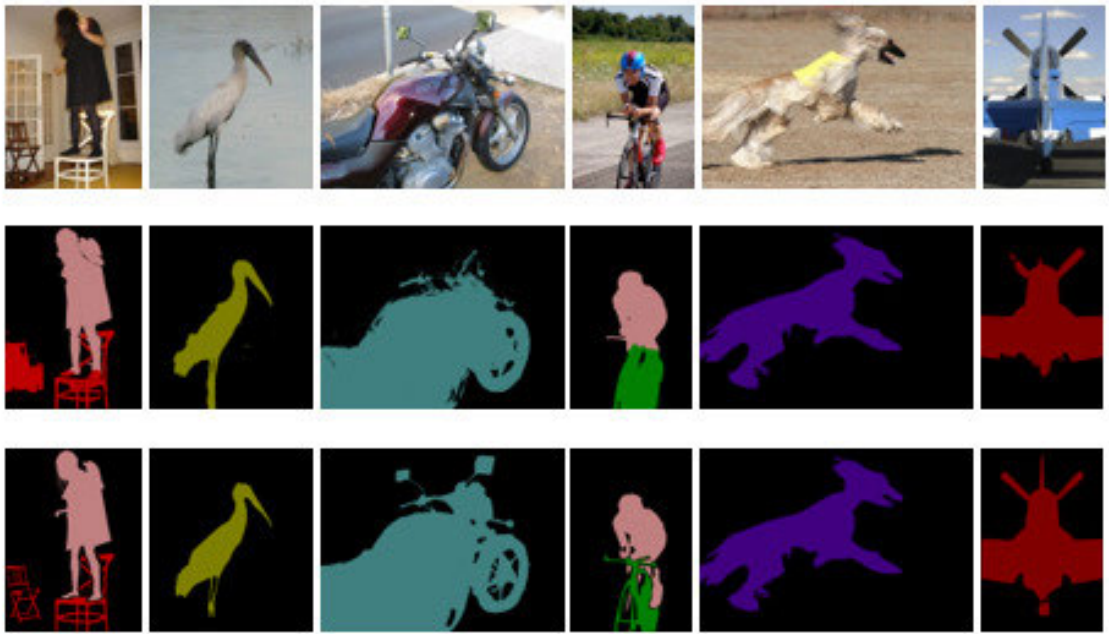

FIGURE 5. The evaluation result in PASCAL VOC 2012 val set: The first row shows the images. The second row shows the predicts of our method(RCDC). The last row shows the ground truth.

The standard *mean Intersection Over Union* (mIOU) is adopted for evaluation on PASCAL VOC val dataset.

As the Table 1 shown, Regional Location Cutting in val set means that use the method of making pseudo-labels in val set. It is obvious that then pseudo-labels are not as good as we need so the dynamic credible regions training correction will be necessary to improve it. The baseline represents that just training the segmentation network with the pseudo-labels generated by Regional Location Cutting, it reaches the performance of 56.15 (mIOU) in PASCAL VOC 2012 val set. Adding *Dynamic Region Loss* into the baseline training, the performance with *Dynamic Region Loss* method has been improved up to 61.05 (mIOU). Retraining the network and adding the fully-connected CRF process to the network, our method finally reaches 63.55 (mIOU). The Fig. 5 shows the predicted results of our method in PASCAL VOC 2012 val set.

C. ANALYSIS OF PARAMETERS

In our final experiment, the structure of the segmentation network is deeplabV3plus-xception. For the hyperparameters, the batch size is 16 and the total iterations is 36k. It has been proved that with those parameters, our method can reach a great performance. We also compared it with other parameters including the network of deeplabV3plus-resnet101, the batch of 4,32, and the iterations of 50k,20k. The results are shown in Table 2

From Table 2, we can see that the xception is a better backbone network for segmentation compare with resnet101. Because some of the pseudo-labels may have a large area of the wrong annotation, the direction of gradient descent with a small batch size of only several bad pseudo-labels may lead the model to a worse result. With the bigger batch size, the model is easy to converge to the local minimum. If we increase the number of iterations, the model may get overfitting and the result may get worse. And the model may get under fitting with fewer iterations.

TABLE 2. The result of different results of using different parameters by evaluated in PASCAL VOC 2012 val set.

network	batch size	iterations	mIOU
deeplabV3plus-xception	16	36k	**63.6**
deeplabV3plus-xception	4	36k	62.7
deeplabV3plus-xception	32	36k	62.5
deeplabV3plus-xception	4	20k	60.6
deeplabV3plus-xception	16	50k	62.3
deeplabV3plus-resnet101	16	36k	61.7

TABLE 3. The comparison with the bounding box level methods evaluated in PASCAL VOC 2012 val set.

Methods	BoxSup	WSSL	SDI	**RCDC(ours)**
Unit (mIOU)	62.0	60.6	69.4	**73.3($\pm$ 0.2)**

D. COMPARISONS WITH STATE-OF-THE-ARTS

Results of other state-of-the-art weakly-supervised semantic segmentation solutions on PASCAL VOC 2012 validation and test dataset are compared with our method(RCDC) in this section. We first compare our bounding box method with other bounding box level semantic segmentation methods: BoxSup [38], WSSL [39], SDI [21]. The Table 3 shows the comparison with other bounding box level methods, our method reaches the performance of 73.3($\pm$0.2)% mIOU:

TABLE 4. The comparison with the image-level methods evaluated in PASCAL VOC 2012 val and test set.

Methods	supervision	val	test
AE-PSL [29]	Image-level	55.0	55.7
SEC [20]	Image-level	50.7	51.7
GAIN [35]	Image-level	55.3	56.8
GAIN [35]	Image-level+ pixel level	60.5	62.1
DCSP-VGG [40]	Image-level	58.6	59.2
DCSP-Res [40]	Image-level	60.8	61.9
SeeNet [41]	Image-level	63.1	62.8
MDC [**?**]	Image-level	60.4	60.8
MCOF-VGG [31]	Image-level	56.2	57.6
MCOF-Res [31]	Image-level	60.3	61.2
DSRG-VGG [32]	Image-level	59.0	60.4
DSRG-Res [32]	Image-level	61.4	63.2
Shen et al. [43]	Image-level+Web images	63.0	63.9
SSNet [33]	Image-level+Saliency	63.3	**64.3**
RCDC(ours)	Image-level	**63.5(±0.2)**	64.1(±0.3)

Then, we compared our method in image-level with other image-level semantic segmentation methods. Some previous state-of-the-arts methods are chosen: SEC, MCOF, DSRG, MDC, *etc*. Table 4 shows the comparison with other image-level methods on PASCAL VOC 2012 val and test set.

In Table 4, the supervision of the SSNet method is not only the image-level labels but the saliency labels which have more information than image-level labels. The method of Shen *et al.* using the supervision of image-level labels of the training set and a large number of web images. The GAIN method has the supervision of image-level training set and 1464 pixel-level labels. All those methods have more supervision information than RCDC but our method RCDC still has a competitive performance.

We also analyze the comparison of the timing cost between manual annotations and RCDC. Generally, the pixel-level annotation may take one or two hours per image for manual annotation while the pixel-level annotations just tale one or two seconds. It is obvious that labeling the segmentation labels takes thousands of times than labeling image-level labels which RCDC is based on. So, RCDC has a huge advantage in timing cost compared with manual annotation.

Our methods reach the best performance compared with other bounding box level methods which have no pre-trained. Compared with the methods only using image-level labels for supervision, our method (RCDC) reaches a competitive performance which achieves 63.5(±)0.2% on PASCAL VOC 2012 val set and 64.1(±0.3)% on the test set. Some methods such as SSNet and the method of Shen et al used not only image-level labels but adding other supervision which will take tremendous amounts of time for manual annotation. Compared with them, our method still has a competitive performance.

V. DISCUSSIONS

Our method RCDC can obtain a high-performance segmentation result with just image-level labels or bounding box labels. It can reduce a lot of manual labeling work.

The two key points of our method are the regional location cutting and the dynamic credible regions training correction. Compared with other weakly-Supervised Semantic Segmentation methods that only use CAM to generate pseudo-labels, our regional location cutting can generate better pseudo-labels for segmentation training. The dynamic credible regions training correction can preserve the regions of great performance during the training of the semantic segmentation network. It can bring a large improvement for better segmentation results. Moreover, our method mainly focuses on making better pseudo-labels and the loss function. It can easily be applied to other better semantic segmentation network or weak methods for better performance.

FIGURE 6. Some wrongly segmentation in PASCAL VOC 2012 val set: The first row shows the images. The second row shows the predicts of our method. The last row shows the ground truth.

We also think about why our methods can reach better performance. After research, we find that a network always tends to learn easy knowledge first and then learn hard knowledge. In our weakly-supervised semantic segmentation method, the right area of pseudo-labels is the easier knowledge while the wrong area is a kind of random noise that can be considered as hard knowledge. During the training, our method can save easy knowledge which is the right knowledge and using this knowledge for further training. With the correction of easy knowledge, hard knowledge can become better and it can lead to better performance.

However, there are also some limitations. The pseudo-labels in our methods are based on the traditional segmentation method GrabCut. The GrabCut sometimes can not get the full foreground of the segmentation target, so, there may be some bad pseudo labels in the training set. It may cause some bad segmentation results. As it is shown in Fig. 6, it is hard for the model to distinguish similar targets like trains and tracks.

VI. CONCLUSION

In this article, we proposed RCDC, a new weakly supervised semantic segmentation method with regional location cutting

and dynamic credible regions correction. It can achieve image segmentation just with image-level labels or bounding box level labels which can greatly reduce the consumption of manual annotation. The regional location cutting can provide the pseudo-labels which can cover about 50% of ground truth. It has about 40% improvement compared with CAM. The dynamic credible regions training correction can preserve great performance during the training to enhance the pseudo-labels. Experiments with PASCAL VOC 2012 dataset revealed that the proposed method achieved the best comprehensive performance compared with other methods. Supervised by bounding box level labels, our method reaches 73.3(±0.2)% mIOU which has the best performance among all the non-pre-trained methods. Supervised by image-level labels, our method has the best comprehensive performance of 63.6% mIOU with minimal supervision.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[4] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 234–241.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[11] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15750–15761, 2017.

[12] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "RIC-unet: An improved neural network based on unet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21420–21428, 2019.

[13] Y. Li and L. Shen, "CC-GAN: A robust transfer-learning framework for HEp-2 specimen image segmentation," *IEEE Access*, vol. 6, pp. 14048–14058, 2018.

[14] Q. Li, A. Arnab, and P. H. Torr, "Weakly and semi-supervised panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 102–118.

[15] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, "Cyclic guidance for weakly supervised joint detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 697–707.

[16] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," 2018, *arXiv:1811.07073*. [Online]. Available: http://arxiv.org/abs/1811.07073

[17] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5208–5217.

[18] M. Tong, W. Lin, X. Huo, Z. Jin, and C. Miao, "A model-free fuzzy adaptive trajectory tracking control algorithm based on dynamic surface control," *Int. J. Adv. Robotic Syst.*, vol. 17, no. 1, 2020, Art. no. 1729881419894417.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?–weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 685–694.

[20] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 695–711.

[21] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 876–885.

[22] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 642–651.

[23] S. Maji, N. K. Vishnoi, and J. Malik, "Biased normalized cuts," in *Proc. CVPR*, Jun. 2011, pp. 2057–2064.

[24] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 328–335.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[28] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 801–818.

[29] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1568–1576.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062

[31] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1354–1362.

[32] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7014–7023.

[33] Z. Yu, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7223–7233.

[34] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut' interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[35] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9215–9223.

[36] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[37] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 30, 2019, doi: 10.1109/TPAMI.2019.2938758.

[38] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.
[39] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a DCNN for semantic image segmentation," 2015, *arXiv:1502.02734*. [Online]. Available: http://arxiv.org/abs/1502.02734
[40] A. Chaudhry, P. K. Dokania, and P. H. S. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," 2017, *arXiv:1707.05821*. [Online]. Available: http://arxiv.org/abs/1707.05821
[41] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Proc. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2018, pp. 549–559.
[42] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for Weakly- and semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.
[43] T. Shen, G. Lin, C. Shen, and I. Reid, "Bootstrapping the performance of webly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1363–1371.

MINGSI TONG (Member, IEEE) received the B.S. degree in mechanical engineering from Northeast Forestry University, China, in 2008, and the M.S. degree in aerospace manufacturing engineering and the Ph.D. degree in mechatronic engineering from the Harbin Institute of Technology, China, in 2011 and 2016, respectively. He was working for the National Institute of Standards and Technology, USA, as a Guest Researcher, from 2012 to 2015. He is currently an Assistant Professor with the School of Mechatronics Engineering, Harbin Institute of Technology. His research interests include machine vision, pattern recognition, and surface metrology.

WENCONG LI received the B.S. degree in automation from the University of Electronic Science and Technology of China, in 2019. He is currently pursuing the M.S. degree in control engineering with the Harbin Institute of Technology, China. His research interests include computer vision, deep learning, and machine learning.

XINYANG REN (Member, IEEE) received the bachelor's degree in automation from Harbin Engineering University, China, in 2016. He is currently pursuing the Ph.D. degree with the Harbin Institute of Technology, Harbin, China. His research interests include computer vision, machine learning, and deep learning.

XINGHU YU received the M.M. degree in osteopathic medicine from Jinzhou Medical University, Jinzhou, China, in 2016. He is currently pursuing the Ph.D. degree in control science and engineering from with the Harbin Institute of Technology, Harbin, China. His research interests include intelligent control and biomedical image processing

WEIYANG LIN received the B.S. and M.S. degrees in mechanical engineering from the Harbin Institute of Technology, China, in 2006 and 2008, respectively, and the Ph.D. degree in mechatronics engineering from the Shenzhen Graduate School, Harbin Institute of Technology, in 2014. He is currently an Associate Professor with the Research Institute of Intelligent Control and Systems, Harbin Institute of Technology. His research interests include parallel manipulators, robotic motion control, and medical robotic design and control.