This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

SAL: Selection and Attention Losses for Weakly Supervised Semantic Segmentation

Lei Zhou, Chen Gong, Zhi Liu, *Senior Member, IEEE*, Keren Fu

***Abstract*—Training a fully supervised semantic segmentation network requires a large amount of expensive pixel-level annotations in manual labor. In this work, we focus on studying the semantic segmentation problem using only image-level supervision. An effective scheme for weakly supervised segmentation is employed to produce the proxy annotations via image tags firstly. Then the segmentation network is retrained on the generated noisy proxy annotations. However, learning from noisy annotations is risky, as proxy annotations of poor quality may deteriorate the performance of the baseline segmentation and classification networks. In order to train the segmentation network using noisy annotations more effectively, two novel loss functions are proposed in this paper, namely, the selection loss and attention loss. Firstly, a selection loss is designed by weighting the proxy annotations based on a coarse-to-fine strategy for evaluating the quality of segmentation masks. Secondly, an attention loss taking the clean image tags as supervision is utilized to correct the classification errors caused by ambiguous pixel-level labels. Finally, we propose an end-to-end semantic segmentation network SAL-Net guided by the above two losses. From the extensive experiments conducted on PASCAL VOC 2012 dataset, SAL-Net reaches state-of-the-art performance with mean IoU (mIoU) as 62.5% and 66.6% on the test set by taking VGG16 network and ResNet101 network as the baselines respectively, which demonstrates the superiority of the proposed algorithm over eight representative weakly supervised segmentation methods. The code and models are available at https://github.com/zmbhou/SALTMM.**

***Index Terms*—Deep Learning, Weakly Supervised Semantic Segmentation, Selection Loss, Attention Loss.**

I. Introduction

Convolution Neural Networks (CNNs) based semantic segmentation or object detection have achieved greater success recently [1], [2], [3], [4], [5], where the goal is to assign a semantic label to a pixel. However, the prediction accuracy of CNNs relies on a large amount of accurate pixel-level annotations, and the collection and annotations of datasets are time-consuming and laborious. Weakly supervised semantic segmentation methods which require less pixel-level annotations are designed to solve the above problem. As to weakly supervised segmentation, there exist several types of weak supervisions: image labels, bounding boxes or scribbles etc. Among these supervisions, image-level labels are the most convenient to be generated by indicating the presence and absence of the classes of interest. There are two main categories of methods for weakly supervised segmentation with image-level supervision. One category focuses on mining discriminative regions iteratively [6], [7], [8], [9]. The other focuses on generating and retraining with high-quality proxy annotations [10], [11], [12], [13], [14]. Accordingly the two most important techniques used in current weakly supervised semantic segmentation methods can be summarized as: (1) mining discriminative foreground and background regions, and (2) retraining the segmentation network with proxy annotations. As to discriminative regions mining, various iterative mining strategies have already been proposed [10], [15], [16], [17], [12], [13], [14], [8], [18]. The procedure of retraining with proxy annotations has also been proven to be effective in boosting the segmentation performance by taking the generated segmentation masks as supervision for final network training [19], [20], [14], [12], [8], [21].

However, retraining from proxy annotations may incur

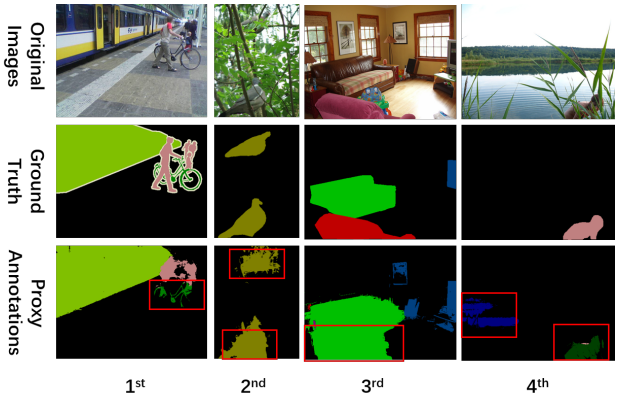

Fig. 1: Illustration of the proxy annotations with poor boundary conditions or severe classification errors. (Objects with poor boundary condition and miss-classified pixels are highlighted by the red rectangles).

This work was supported by the National Natural Science Foundation of China under Grants 61906121, 61973162 and 61771301, NSF of Jiangsu Province (No: BK20171430), the Fundamental Research Funds for the Central Universities (No: 30918011319), the "Summit of the Six Top Talents" Program (No:DZXX-027), the "Young Elite Scientists Sponsorship Program" by Jiangsu Province, and the "Young Elite Scientists Sponsorship Program" by CAST (No: 2018QNRC001) (Corresponding author: Lei Zhou).

L. Zhou is with the School of Medical Instrument and Food Engineering, and is also with Shanghai Engineering Research Center of Assistive Devices, University of Shanghai for Science and Technology, Shanghai, China (email:davidzhou@usst.edu.cn).

C. Gong is with the PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, and is also with the Department of Computing, Hong Kong Polytechnic University (e-mail: chen.gong@njust.edu.cn).

Z. Liu is with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: liuzhisjtu@163.com).

KR. Fu is with the College of Computer Science, Sichuan University, Chengdu, China (e-mail: fkrsuper@scu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

errors if they are of poor quality. When the noisy proxy annotations are used, the training procedure will suffer from two potential risks: (1) noisy proxy annotations with bad boundary condition (see the 1st and 2nd examples in Fig. 1) may deteriorate the segmentation performance sharply and (2) ambiguous categories may be wrongly classified due to the confused pixel-level labels contained in proxy annotations (see the 3rd and 4th examples in Fig. 1). In the first case, the proxy annotations corresponding to complex images (such as images with small objects or clutter background) are of poor boundary accuracy and this kind of proxy annotations can deteriorate the segmentation performance sharply according to our observation. In the second case, the classification ambiguities between objects with similar categories worsen due to the confusing pixel-level labels.

Based on the above observations, we argue that not all proxy annotations are suitable to participate in the retraining process, especially when they are of poor boundary condition or with serious classification errors. Therefore, we evaluate the quality of the segmentation masks and correct the miss-classification adaptively. This modification to the traditional scheme of retraining is very critical, because in this modification the annotations with high quality are selected and classification errors can be refined, so that the segmentation network can be optimized more effectively.

By taking advantage of these psychological opinions, a segmentation framework with two novel losses (selection loss and attention loss) is proposed in this paper (displayed in Fig. 2). The selection loss is defined to assign low confidence to the bad proxy samples by evaluating the quality of masks in a coarse-to-fine way. Then the proxy annotations of high confidence are used to retrain the segmentation network. The attention loss is built on a classification structure, then the segmentation network can be trained with clean image tags by adjusting the classification ambiguity caused by the noisy samples adaptively. Furthermore, the attention weights are used to refine the segmentation probability in the training process as well. Finally, the whole framework is optimized by the selection loss and attention loss jointly in an end-to-end way.

Overall, our major contributions can be summarized as follows:

- We introduce a coarse-to-fine mask scoring strategy for evaluating the quality of proxy annotations and then a selection loss is proposed for optimizing the weakly supervised semantic segmentation network with high-confidence annotations.
- We propose a classification subnetwork with hybrid dilation convolutions guided attention loss to adjust the classification errors by learning from clean image tags and interacting with segmentation network adaptively.
- We employ a simple and effective training protocol based on selection loss and attention loss, which is different from most existing methods of weakly-supervised semantic segmentation.
- Detailed ablation experiments have been conducted to verify the effectiveness of the proposed losses. Our work obtains the state-of-the-art weakly supervised semantic segmentation performance on the PASCAL VOC 2012 segmentation benchmark. The mIoU of our method are 62.5% and 66.6% on the test set using VGG16 and Resnet101 as the baselines respectively.

II. Related Work

Weakly supervised approaches have been widely studied for semantic segmentation. Various weak supervisions such as bounding boxes [22], scribbles [23] or image-level tags [10], [11], [24], [17], [12], [13] have been used to improve the segmentation efficiency. In this paper, we focus on the image-level supervised framework. We will briefly review the existing approaches from the following two aspects:

A. Mining Discriminative Regions in Weakly Supervised Segmentation

Mining discriminative regions is the most important technique in training semantic segmentation model with image-level supervisions. There are mainly four strategies for mining regions: (1) localization with classification DCNNs, (2) mining with saliency cues, (3) iterative erasing strategy and (4) hybrid training strategy. The first kind of strategies identify the discriminative regions respecting to each individual class based on image classification DCNNs. Zhou et al. [25] proposes a technique called Class Activation Mapping (CAM) for identifying discriminative regions by replacing fully-connected layers in image classification CNNs with convolutional layers and global average pooling layer. CAM [25] is the most widely used technique in weakly-supervised semantic segmentation for generating pseudo-annotations [10], [26], [27]. Grad-CAM [28] is a strictly generalization of CAM without the need of modifying DCNN structure. Different from the previous mentioned strategies, Zhang et al. [29] proposes Excitation Backprop to back-propagate in the network hierarchy to identify the discriminative regions. The second category of strategies take the saliency information as guidance. In [24], a saliency guided weakly-supervised semantic segmentation framework is presented to evaluate different fusion strategies comprehensively. Moreover, the saliency cues have also been used to refine the discriminative regions to boost the segmentation performance [11], [7], [30]. The third category of strategies focus on iterative erasing, in which the class-specific discriminative regions are discovered in a hide and seek manner. For example, in [26], an adversarial erasing approach is proposed to effectively adapt a classification network to progressively discovering and expanding object discriminative regions. In [7], two self-erasing strategies are designed to prohibit the attention regions from spreading to unexpected regions given the roughly accurate background priors. In [30], a generic classification network equipped with convolutional blocks of different dilated rates is proposed to generate dense localization maps. In [8], many different localization maps are generated from a single image by choosing features at random during both training and inference, then those maps are aggregated into a single localization map.

Different from the above three strategies, the hybrid training

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

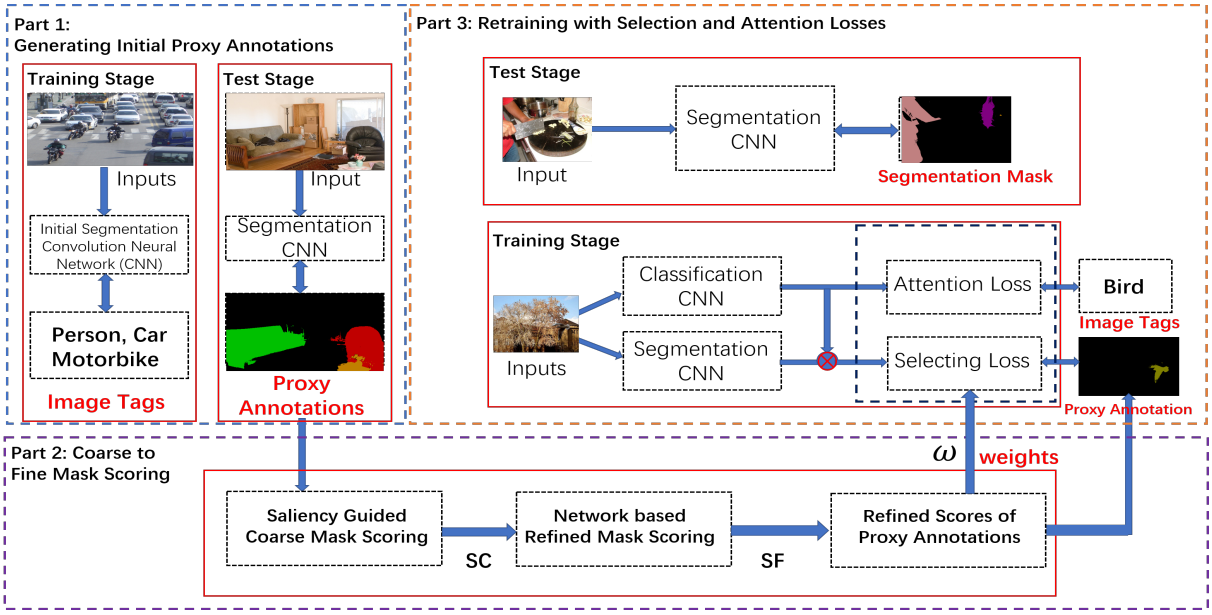

Fig. 2: The flowchart of the proposed weakly supervised semantic segmentation framework. The whole framework consists of three parts: generating initial proxy annotations, coarse-to-fine mask scoring, and retraining with selection and attention losses.

strategy focuses on generating high-quality proxy annotations iteratively. For example, in [10], a "seed, expand and constrain" (SEC) framework is proposed using only image-level labels where localization cues from classification networks are used to find the objects. In [11], a class-specific pixel discovery method for weakly-supervised semantic segmentation method is designed. The initial localization cues are combined with the saliency cues to generate the proxy annotations. In [17], an iterative bottom-up and top-down framework is presented which tolerates inaccurate initial localizations by iteratively mining common object features from object seeds. In [12], a deep seeded region growing (DSRG) training approach is designed which gradually improved the quality and extent of mined object regions. DSRG also reveals that the retraining procedure with proxy annotations refined by Condition Random Filed (CRF)[31] can boost the segmentation performance significantly. In [13], the authors present a novel framework based on AffinityNet which is used for generating accurate segmentation labels of training images given their image-level class labels only. The refined proxy annotations generated by the AffinityNet demonstrate higher quality for training the segmentation model.

B. Learning From Noisy Proxy Annotations

Learning from noisy samples is a mechanism to train the segmentation model with noisy image tags or proxy annotations. Because handling noisy samples is an important procedure in the task of weakly supervised semantic segmentation, until now many methods have been explored from different perspectives. As to learning from noisy image tags, a ℓ_1 optimized based sparse learning model is formulated to identify and correct the superpixel noisy labels in [32]. In [33], the weakly supervised segmentation problem is transformed into a large-scale sparse learning problem by learning the data manifolds. As to learning from proxy annotations, many methods have been proposed. In [34], a method is proposed to sanitize the annotations and measure their reliability, so as to alleviate the side effects introduced by noisy and incomplete annotations. In [35], a two-stream mutual attention network is presented to discover incorrect labels and to weaken the influence of these incorrect labels during the parameter updating process. In [36], a Filling Rate guided adaptive loss (FR-loss) is suggested to help the model ignore the wrongly labeled pixels in proposals. FR-loss can adjust the model learning with global statistical information.

Different from the above methods, more solutions have been designed from the perspective of high-quality annotations selection [37], [19], [20], [14]. For example, in [37], Wei et al. proposes a simple to complex framework where a network is first trained using simple images (single object category) followed by training over complex ones (multiple objects). In [14], a novel bi-directional transfer learning framework is designed to generate high quality masks for the training images. However, few of these methods is able to solve the problem of learning from noisy pixel-level annotations for semantic segmentation effectively.

In order to improve the segmentation performance by learning noisy pixel-level annotations more effectively, our method focuses on designing novel mask scoring mechanisms for selecting high-quality annotations, and then two novel loss functions are proposed to retrain the segmentation network.

III. The Proposed Method

As shown in Fig. 2, the proposed framework is comprised of three principal components: (1) Generating initial proxy annotations by taking image tags as the supervision; (2) Selecting high-quality annotations by a coarse-to-fine mask

scoring strategy; (3) Retraining the segmentation network with selection and attention losses jointly. The details will be introduced in the following subsections.

A. Generating Initial Proxy Annotations

In order to learn from the image tags, we apply the SEC [10] and DSRG [12] frameworks for learning from the image-level labels. Let $T = (I_n, Y_n)^N$ be the samples in the set of training dataset which consists of N images. The image I_n is annoated by image-level labels $Y_n \in \{0,1\}^C$ where C is the number of classes. The semantic segmentation model is designed using deep neural networks $Z(I;\Theta)$ with parameters Θ and Z represents the category probabilities. The initial SEC model is trained by three losses:

$$L_{weak} = \sum_{p=1}^{N} L_{seed}(Z(I_p;\Theta), Y_p) + L_{expand}(Z(I_p;\Theta), Y_p) + L_{constraint}(Z(I_p;\Theta), Y_p) \tag{1}$$

L_{seed} is supervised by the localization cues learned from the Class Activation Mapping (CAM) [25]. Given the image-level labels, the CAMs method is applied to localize the regions of foreground classes. In the procedures of CAMs, the classification network is initialized with VGG-16 network. Then the global average pooling (GAP) is applied on conv7 layer to aggregate the features. Finally the generated feature tensors are classified using a fully-connected layer and the heatmap corresponding to each category is generated via classification. It is obvious that the discriminative regions can be generated by applying a hard threshold to the heatmap. Hence the role of L_{expand} is to aggregate the heat maps to be consistent with image-level labels by applying a global weighted rank pooling (GWRP) operation. $L_{constraint}$ is designed to enforce the boundary constraint by utilizing the condition random field model as a postprocessing procedure. Then the seed region growing (SRG) strategy proposed in [12] is applied to improve the segmentation performance further. In the growth process of SRG, the image is segmented into regions with the property that each connected component of a region contains exactly one of the initial seeds. The label for each pixel is updated iteratively by the SRG strategy simultaneously.

Once the segmentation network $Z(I;\Theta)$ was trained. The images in the training set are segmented via the initial segmentation model and proxy pixel-level annotations $M = \{G_1, .., G_N\}$ can be generated. In a typical retraining procedure, the segmentation network is optimized with all the proxy annotations using the cross-entropy loss:

$$L_{retrain} = -\sum_{p} \sum_{(x,y)\in I_p} \sum_{c} G_p^c(x,y) log(Z^c(I_p(x,y);\Theta)), \tag{2}$$

where $G_p^c(x,y) = 1$ if the label at pixel (x,y) is c, otherwise $G_p^c(x,y) = 0$. $Z^c(I_p(x,y);\Theta)$ is the class specific probability of the segmentation network.

Retraining the segmentation network with proxy annotations M has been proven to be effective in boosting segmentation performance. However, the overall quality of the proxy annotations is noisy. The principal risk of retraining with proxy GTs is that the segmentation performance may be deteriorated when some proxy annotations are of poor quality. In order to train the semantic segmentation network from noisy and weak annotations more effectively, a selection loss and an attention loss are proposed to train the network and we will introduce the details in the following subsections.

B. Selection loss

The potential risks of annotations in M can be summarized as two folds: (1) Coarse object boundaries and (2) Pixel-level miss-classification. The selection loss will focus on distinguishing the proxy annotations with poor quality. In order to select high-quality masks, a coarse-to-fine strategy for evaluating the quality of masks in M is designed firstly. The overall pipeline of the mask scoring strategy is illustrated in Fig. 3 which consists of two steps: saliency guided coarse mask scoring and network based fine mask scoring. The coarse mask scoring strategy can be divided into two parts: generating saliency guided reference masks and defining rules for calculating scores. The fine mask scoring strategy contains three steps: defining the network architecture, selecting samples for network training and network parameters optimization.

1) Coarse Mask Scoring Strategy: The coarse scores are calculated via analyzing the boundary accuracy and the classification accuracy. For each segmentation mask, the object boundary accuracy AC and class distribution CD are calculated. When the precise annotations do not exist, we take the saliency cues as coarse guidance. Firstly, the saliency cues are used to generate the reference masks H and they will be compared with the binary masks B of proxy annotations to measure the boundary accuracy. Then the classification errors will be measured by a simple class distribution measure.

Generating Reference Masks H: The widely used saliency method such as DSS [38] is applied to generate the reference masks. Firstly the saliency cues are produced for the images in the training set. Then the fully connected CRF is applied as the post-processing procedure to refine the saliency masks and the reference mask for image i is represented as H_i (The reference masks are illustrated in Fig. 4). For images with multiple objects, a scheme with masks erasing similar to the one used in [11] is designed to highlight multiple salient objects iteratively.

Measuring Object Boundary Accuracy AC: Given the saliency guided reference mask H_i, the binary segmentation mask B_i of proxy annotation G_i is evaluated by the widely used criteria F-measure [39] and the simple objectness score F [40]. Generally speaking, ground truth (GT) is required for calculating the F-measure. However, the precise annotations don't exist in the setting of weakly supervised segmentation, hence the reference masks H_i is taken as the proxy GT to measure the quality of the proxy annotation G_i in a coarse way. Then the object boundary accuracy AC_i is represented as:

$$\begin{aligned} AC_i &= \frac{2 \times prec(H_i,B_i) \times rec(H_i,B_i)}{prec(H_i,B_i) + rec(H_i,B_i)} \times F(B_i)^b, \\ prec(H_i,B_i) &= \frac{|H_i \bigcap B_i|}{|H_i|}, rec(H_i,B_i) = \frac{|H_i \bigcap B_i|}{|H_i \bigcup B_i|} \end{aligned} \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

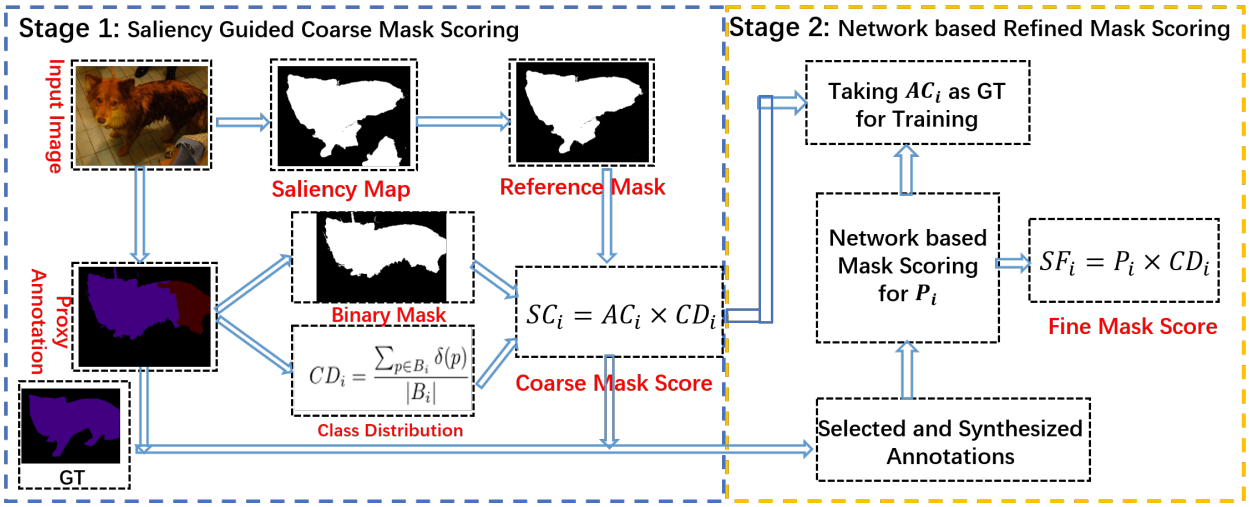

Fig. 3: The flowchart of the proposed coarse-to-fine mask scoring strategy. It consists of two parts: saliency guided coarse mask scoring and network based refined mask scoring.

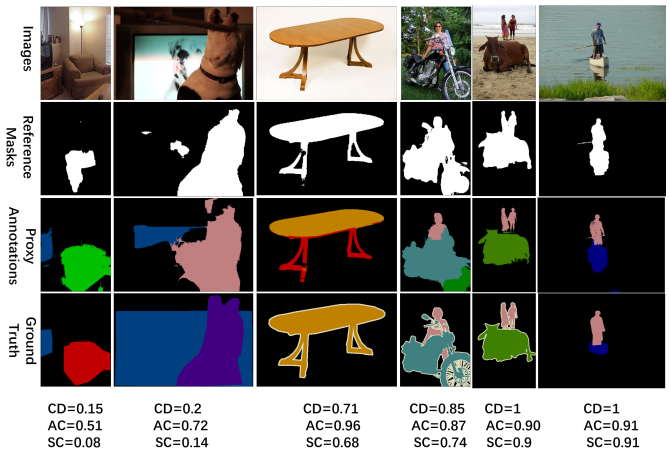

Fig. 4: Illustration of the coarse mask scoring strategy. AC is the boundary accuracy as defined in Eq. (3). CD is the class distribution as defined in Eq. (4). SC is the coarse score as defined in Eq. (5)

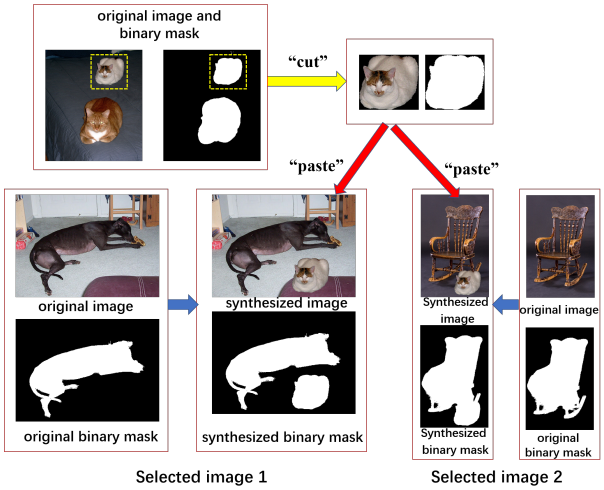

Fig. 5: The flowchart for synthesizing new samples via "cut" and "paste". Two images are selected to display the procedure for synthesizing the samples with multiple instances.

where $prec$ stands for the precision, rec represents the recall and b is the weight cofficient of objectness score.

Measuring the confidence of Class Distribution CD: In the setting of weakly supervised segmentation, it's hard to measure the accurate classification accuracy without precise pixel-level annotations. In the proposed framework, a simple class distribution confidence is designed to measure the classification errors statistically:

$$CD_i = \frac{\sum_{p \in B_i} \delta(p)}{|B_i|}, \delta(p) = 1 \ if \ L_p \in Y_i. \quad (4)$$

where L_p is the label for pixel p, Y_i is set of tags for image i and $|B_i|$ is the number of pixels that belong to the binary objects B_i. The intuition behind the above equation is that if the pixels are classified with the labels that are not contained in Y_i, the confidence of class distribution will decrease. Finally, the coarse scores of masks are calculated by combining the boundary accuracy and class distributions:

$$SC_i = AC_i \times CD_i, \quad (5)$$

The reference masks, the class distribution probabilities and the boundary accuracies are as illustrated in Fig. 4. It is shown that the class-distribution probability is effective in distinguishing annotations with serious miss-classification. However, we note that the object boundary accuracy depends on the quality of the reference masks heavily and the coarse scoring strategy may not work well when the saliency maps fail to capture the semantic objects (see the examples in 1st and 2nd column in Fig. 4). In order to design a more robust scoring strategy, a network based fine mask scoring strategy will be introduced in the next subsection.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

2) Network based Fine Mask Scoring: Recently, many work [41], [42] has revealed that predicting the segmentation accuracy is helpful in boosting the performance of high-level tasks such as object detection or instance segmentation. Motivated by these works, a predictor network is designed for robust mask scoring. The detailed network structure is illustrated in Table I, a nine layers network with convolutions, residual blocks [43] and fully connected layers is proposed to predict the mask evaluation score, specially the GDN [44] activation function is applied to improve the prediction performance. GDN can serves as an attention mechanism and it has been proven effective in the tasks such as predicting image quality[45]. Given the network architecture, the original image are contacted with the binary segmentation mask to construct the input with 4 channels and the output P is taken as the prediction score.

Prediction Network Training: Our methods formulate the mask scores estimation as a regression task. For training the prediction network, we use the selected proxy annotations as training samples. For generating the regression target for each training sample, we firstly get the predicted mask of the target class and binarize the predicted mask. Then we use the scores AC between the binary mask and its reference defined in Eq. (3) as the prediction target. In order to adjust the wrongly predicted mask scores of AC, the samples with high-confident scores are selected to construct the training set. For example, a high threshold AC $\geq$0.8 is set to select around 4k high-quality training images to construct the positive samples. Then a lower threshold such as AC $\leq$ 0.3 is used to select around 1k images as the negative samples. In order to generate the samples with AC among 0.3 to 0.8, the high-quality annotations are eroded or erased to generate another $3k$ samples and the corresponding AC is calculated via comparing with the original reference masks. Moreover, in order to enrich the training samples for the images with multiple instances, the schemes such as "cut" or "paste" proposed in [46] are applied to synthesize more training samples (the whole procedure is as illustrated in Fig. 5). The foreground masks of images with AC $\geq$ 0.9 (around 1k) are cut and selected, and then they are pasted on the background regions of selected images chosen from the samples with AC$\geq$0.3. Therefore, another 2k samples with multiple classes can be generated. Finally, the training set is augmented to an amount of 10k. For each sample in the training set, the score AC predicted by the coarse selection or sample augmentation strategy is treated as the ground truth. As to the loss functions for network training, the ℓ_2-loss as the empirical loss function is utilized in the training procedure:

$$\ell_2(P_i, AC_i) = \|P_i - AC_i\|_2 = (P_i - AC_i)^2, \qquad (6)$$

where P_i is the output of prediction network and AC_i is the coarse boundary accuracy scores of sample i in the training set. We have also tried ℓ_1-normal as the loss, but observe worse results. Once the prediction network is trained, the fine evaluation score SF is formulated as:

$$SF_i = P_i \times CD_i. \qquad (7)$$

As shown in Fig. 6, the role of network based mask scoring

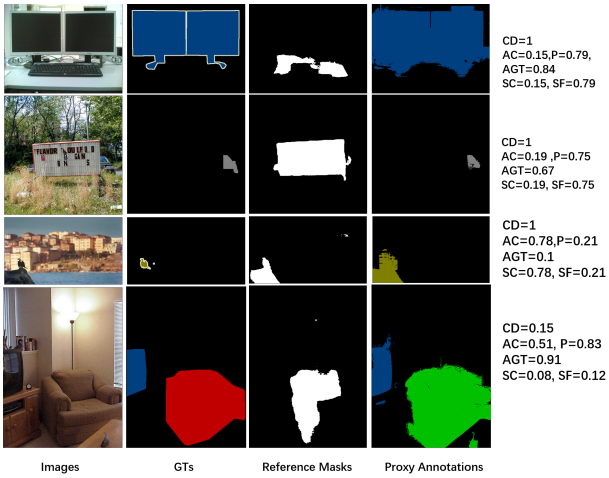

Fig. 6: The illustration on the role of network based mask scoring. AC is defined in Eq. (3), CD is defined in Eq. (4), P is the output of the prediction network. SC is defined in Eq. (5) and SF is defined in Eq. (7). AGT is the boundary accuracy calculated by comparing with the pixel-level ground truth.

is illustrated. When the reference masks fail to fit the semantic objects well, the boundary accuracy scores will be incorrect even if the extracted objects are acceptable (see the 1st, 2nd and 4th examples). On the contrary, the coarse scores may be miss-leading when the reference masks fit the objects in the proxy annotations well but are of low precision compared with the ground truth (see the 3rd example). It's obvious that the proposed network based scoring strategy can generate more reliable quality evaluation scores. Take the 3rd example in Fig. 6 for example, the bird mask is over-segmented. However, it fits well with the reference masks and the simple boundary accuracy is calculated as 0.78. The prediction network can generate a more reliable score as 0.21, which is closer to the ground truth score 0.1.

3) Mask Scoring Guided Selection loss: Note that the wrongly labeled regions of the pixel-level proposals have negative effects on model training, recognizing the negative regions will be helpful. A possible solution is to ignore the pixels with small confident values in the score map, which may be the wrongly labeled pixels. For weakly supervised model, there are no guaranteed pixel-level annotations like the fully supervised model. Thus it is hard to determine how much percentage of pixels to be ignored. Furthermore, we intuitively find that the proxy annotations of low scores can worsen the performance. Then we introduce a selection loss (Sel loss) by setting a weight for each annotation. The masks with high scores are selected as the final proxy annotations for fully supervised learning. Similar to the initial segmentation network training, the selection loss (Sel loss) is formulated as:

$$L_{sel} = -\sum_{i=1}^{N} \sum_{(x,y)\in G_i} \sum_{c=1}^{C} w_i G_i^c(x,y) log Z^c(I_i(x,y);\Theta), \qquad (8)$$

TABLE I: The network architecture for fine mask scoring. Channels denote the parameters of a module as input channel × output channel. Conv stands for the convolution operation and ResB represents the residual block defined in [43].

Methods	**layer 1**	**layer 2**	**layer 3**	**layer 4**	**layer 5**	**layer 6**	**layer 7**	**layer 8**	**layer 9**
Module	3 × 3 Conv	ResB	3 × 3 Conv	ResB	3 × 3 Conv	ResB	3 × 3 Conv	fc	fc
Stride	1	1	2	1	2	1	2	1	1
Channels	4 × 128	128 × 192	192× 192	192 × 192	192× 192	192 × 192	192 × 192	192 × 128	128 × 1
Activations	ReLU	GDN	ReLU	GDN	ReLU	GDN	ReLU	GDN	ReLU
Width	256 × 256	256 × 256	128 × 128	128 × 128	64 × 64	64 × 64	32 × 32	128 × 1	1 × 1

where G_i stands for the i-th proxy annotation, (x, y) represents a pixel location, $Z^l_{xy}(i)$ is defined as the segmentation probability, and $G^c_i(x, y)$ is the ground truth indicator of proxy annotation. w_i is the weight which indicates the confidence of sample i in the training procedure. Then w_i is defined as:

$$w_i = \begin{cases} 1, & \text{if } SF_i \geq \tau; \\ 0, & \text{if } SF_i < \tau. \end{cases} \tag{9}$$

where τ is the hard threshold for masks selecting, which is set by an evaluation procedure for different networks in the experiment section.

C. Attention Loss

Another potential risk when training with noisy proxy annotations is that the pixel-level classification errors may get worse. Even if the noisy annotations with poor quality are filtered, the miss-labeled pixels in selected annotations can also make the classification performance worsen. Take the second image in Fig. 4 for example, the pixels belonging to category dog may be miss-classified as person, even if they are of high boundary accuracy. In the field of weakly supervised segmentation, only the image tags are clean and reliable. In order to correct the classification errors adaptively, a classification subnetwork is constructed to design an attention loss which also can be interacted with the segmentation branch to refine the segmentation probabilities.

1) The Classification Subnetwork: In our formation, the classification network for attention loss is built on a hybrid dilation structure, each branch consists of convolutions with different dilation convolution. According to the validation experiment, the final classification network consists of four branches with the dilation rates as $r = 1, 3, 6, 9$ respectively. The detailed classification network architecture for attention loss and the interaction between the segmentation network is shown in Fig. 7. The segmentation network and the classification network share the same baseline network and the segmentation network consists of a baseline network and the LargeFOV or atrous spatial pyramid pooling (ASSP) module [47]. Assume the corresponding outputs of four branches are represented as d_1, d_2, d_3 and d_4. Then the output of the classification network is formulated as a weighted sum:

$$d_s = d_1 + d_2 + d_3 + d_4. \tag{10}$$

Furthermore the outputs of the classification branch are interacted with the segmentation branch via dot product to adjust the segmentation probabilities adaptively. The operation is as shown in Fig. 7. Assume the output of segmentation network is represented as Z, then the attention guided probability $\hat{Z}$ can be formulated as:

$$\hat{Z} = Z \bigotimes d_s, \tag{11}$$

where $\bigotimes$ stands for the channel-wise multiplication.

2) Formation of the Attention Loss: Based on the output of classification branch d_s and the clean image tags Y, the attention loss (Atten loss) is formulated as:

$$L_{atten} = -\sum_{i=1}^{N}\sum_{c=1}^{C} Y_i(c) \times log(d^c_s(i)), \tag{12}$$

where N is the number of training samples and $d^c_s(i)$ is the classification score for sample i corresponding to class c. $Y_i(c) = 1$ is the set of image tag contains class c, otherwise $Y_i(c) = 0$.

D. Network Training with Selection loss and Attention Loss

1) The Training Protocol: The training protocol can be summarized as four steps. Step 1: The initial segmentation network is trained with the image tags, and the proxy annotations are generated. Step 2: Defining the selection loss based on a coarse-to-fine mask scoring strategy. Step 3: Defining the attention loss based on a classification network. Step 4: The final segmentation network is trained with the selection loss and the attention loss jointly.

2) The Loss Functions: For constructing the selection loss, the attention weight is enforced on the segmentation probabilities Z as described in Eq. (11) to generate the refined probability $\hat{Z}$ (shown in Fig. 7). Then the selection loss is reformulated as:

$$L_{sel} = -\sum_{i=1}^{N}\sum_{(x,y)\in M_i}\sum_{c=1}^{C} w_i G^c_i(x, y) log\hat{Z}^c_{xy}(i), \tag{13}$$

It is noted that two types of selection losses: coarse selection loss and refined selection loss are defined. The coarse selection loss is constructed by calculating the w_i in Eq. (9) via the coarse mask scoring rule while the refined selection loss is based on the w_i calculated via fine mask scoring rule. Finally, the combined loss function is formulated as:

$$L = L_{sel} + \lambda L_{atten}, \tag{14}$$

where λ controls the weight of two losses. L_{sel} is defined in Eq. (8) and L_{atten} is defined in Eq. (12). Finally , the combined loss functions are used to optimize the whole semantic segmentation network in an end-to-end way.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

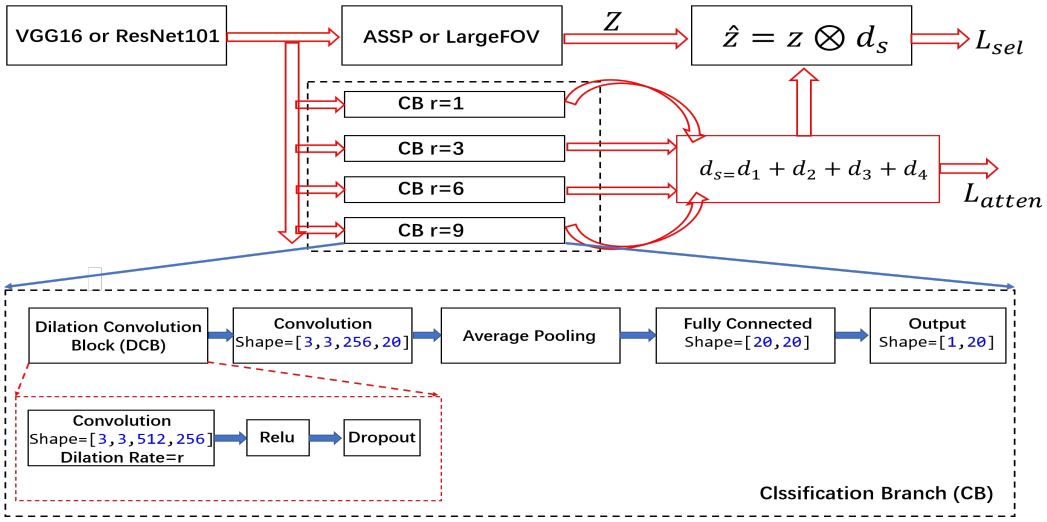

Fig. 7: Illustration of the structure of selection loss and the classification network with four branches for attention loss. r stands for the dilation rate.

IV. Experimental Results

In this section, we first describe the experiment setups, and then report the performance of our approach and compare it with previous state-of-the-art methods. At last, we conduct a series of experiments to demonstrate the impact of each component of our proposed approach on the performance.

A. Dataset and Experimental Setup

This section demonstrates the effectiveness of our approach with comparisons to current state-of-the-art weakly supervised semantic segmentation methods on the PASCAL VOC 2012 segmentation benchmark [48]. As to the performance metric, we adopt the Intersection-over-Union (IoU) between ground truth and predicted segmentations. We evaluate our framework on the challenging PASCAL VOC12 segmentation benchmark dataset [48], which contains 20 foreground object categories and one background category. The original dataset contains 1,464 training images. Following common practice, we augment the dataset with the extra annotations provided by [49]. This gives us a total of 10,582 training images. The validation and test sets contain 1,449 and 1,456 images, respectively. No additional data is being used in the entire train/test pipeline. In our experiments, we only utilize image-level class labels of the training images. We use the val images to evaluate our method. As for the evaluation measure for segmentation performance, we use the standard PASCAL VOC 2012 segmentation metric: mean intersection-over-union (mIoU). We implement two segmentation frameworks: SAL-Net-VGG16 built on Deeplab-ASSP [47] and VGG16 [50], and SAL-Net-ResNet101 built on DeepLab-ASSP [47] and Resnet101 [43]. Our approach is implemented based on tensorflow. In the selection process, we set $b = 0.1$ in Eq. (3) and λ=2 in Eq. (13) experimentally. In the training process, SGD with mini-batch is used for training the classification and segmentation networks. We use the momentum of 0.9, a weight decay of 0.0005, a dropout rate as 0.5 and the batch size is 7. The initial learning rate is $blr = 1 \times 10^{-3}$ for SAL-Net-VGG16 and $blr = 1 \times 10^{-4}$ for SAL-Net-ResNet101. The base learning rate is decreased by the poly learning policy:

$$clr = (1 - \frac{iter}{maxiter})^{power}.blr, \tag{15}$$

where $power$ = 0.9, $maxiter$=27000 for VGG16 based network and $maxiter$=15000 for ResNet101 based network. $iter$ denotes current iteration number and clr is the current learning rate. In the test phase, the learned segmentation network is applied to produce the probability map for each testing image. Then, we upscale the predicted probability map to match the size of the input image, and then apply multi-scale (MS) fusion and a fully-connected CRF [47] to refine the segmentation results. For the mask scoring network, we set the initial learning rate as 1×10^{-5}, the batch size as 8 and the epoch number as 50.

B. Comparison with State-of-the-art Methods

The proposed method is compared with previous state-of-the-art image-level supervised semantic segmentation methods, SEC [51], Pixel Affinity [13], BOOTSTRAPPING [14], DSRG [12], MDC [30], SeeNet [7] and FickleNet [8] built on VGG16 or ResNet101. The two proposed implementations SAL-Net-VGG16 and SAL-Net-ResNet101 are evaluated, and the results on PASCAL VOC validation and test datasets are summarized in Table II and Table III, respectively. When using VGG-16 as the basic network, SAL-Net-VGG16 achieves mean IoU of 61.3 and 62.5[1] for val and test sets. When the more powerful ResNet101 is used, SAL-Net-ResNet101 achieves mean IoU of 66.1 and 66.6[2] for val and test sets, outperforming the second best method FickleNet by 1.2% and 1.3%, respectively. Compared with the latest research FickleNet, our framework has been designed from a new viewpoint and our segmentation performance has been improved. By selecting high-quality annotations and utilizing

[1]http://host.robots.ox.ac.uk:8080/anonymous/KEOAJP.html
[2]http://host.robots.ox.ac.uk:8080/anonymous/L3XWLR.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

TABLE II: Overall accuracy on PASCAL VOC 2012 val dataset. The IoUs of twenty categories and the mean IoU are presented.

Method	bk	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
SEC [51]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
Decouple-VGG16 [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.4
Pixel Affinity (VGG16) [13]	87.2	57.4	25.6	69.8	45.7	53.3	76.6	70.4	74.1	28.3	63.2	44.8	75.6	66.1	65.1	71.1	40.5	66.7	37.2	58.4	49.1	58.4
BOOTSTRAPPING (VGG16) [14]	85.0	74.4	24.9	76.2	20.7	58.2	82.3	73.6	81.0	25.9	71.3	37.4	71.8	69.6	70.3	71.0	44.1	73.8	34.1	48.4	40.0	58.8
DSRG (VGG16) [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59.0
Revisit (VGG16) [30]	89.5	85.6	34.6	75.8	61.9	65.8	67.1	73.3	80.2	15.1	69.9	8.1	75.0	68.4	70.9	71.5	32.6	74.9	24.8	73.2	50.8	60.4
SeeNet (VGG16)[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.1
FickleNet (VGG16) [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.2
SAL-Net-VGG16	89.1	78.2	33.7	72.5	53.6	70.2	78.6	70.5	79.6	22.5	71.7	24.9	74.2	68.8	69.5	68.3	37	76.7	28.5	71.7	51.8	**61.3**
Decouple-ResNet101 [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	58.2
Pixel Affinity (ResNet38) [13]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
BOOTSTRAPPING (ResNet50) [14]	86.8	71.2	32.4	77.0	24.4	69.8	85.3	71.9	86.5	27.6	78.9	40.7	78.5	79.1	72.7	73.1	49.6	74.8	36.1	48.1	59.2	63.0
DSRG (ResNet101) [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
SeeNet (ResNet101)[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.1
FickleNet (ResNet101) [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.9
SAL-Net-ResNet101 (ours)	90.2	75.5	31.2	72.8	56.6	71	88	78.3	88.3	26.7	76.9	31.9	79.9	75.2	70.8	76.5	54.1	77.9	35.9	64.8	64.8	**66.1**

clean image tags adaptively, SAL-Net-VGG16 further boosts the segmentation performance by 2.3% on the validation set and by 1.9% on the test set when compared with DSRG, while FickleNet with VGG16 boosts the segmentation performance by 2.2% on the validation set and by 1.5% on the test set when compared with DSRG. Greater improvement can be observed for SAL-Net-ResNet101 when a more powerful backbone network is utilized. Therefore, the segmentation performance can be boosted from 61.4 to 66.1 on the validation set and 4.7% performance gain can be achieved, while FickleNet only achieves a mIoU gain of 3.5%.

Fig. 8 shows examples from the validation set. The segmented masks generated by state-of-the-art methods such as DSRG [12], SeeNet [7] may suffer from poor boundary condition and many discriminative regions are ignored. In contrast, the proposed algorithm can select more high-quality annotations for network training and apply the attention mechanism for segmentation probability refinement. Thus, more discriminative regions with high classification confidence can be discovered as shown in the results of SAL-Net-VGG16.

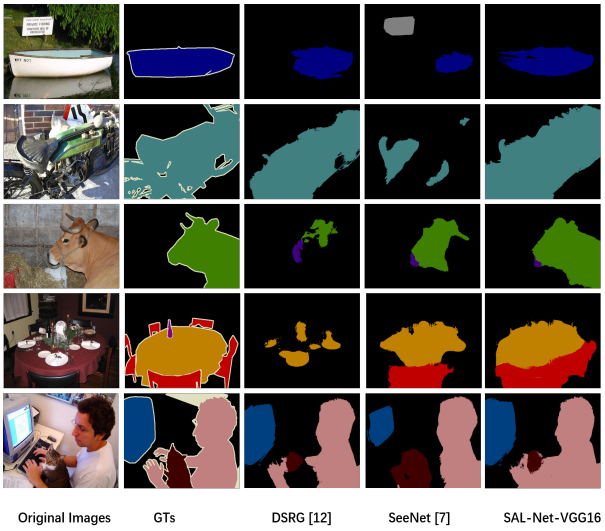

Fig. 8: Predictions on PASCAL VOC 2012 validation set for comparing with different methods such as DSRG [12] and SeeNet [7].

C. Analysis on the Mask Scoring Strategy

The impact of the threshold τ corresponding to the coarse and refined mask scoring strategies on the segmentation performance is analysed. Firstly, the thresholds for the coarse mask scores SC defined in Eq. (5) and validation results are reported in Table IV. For VGG16 based model, we find that the segmentation network is sensitive to the amount of training samples and the segmentation performance will drop sharply when the number of training images decreases. According to the cross validation, the setting SAL-Net-VGG16 with $\tau = 0.55$ achieves the highest mIoU as 57.6, which is selected as the setting for the coarse selection loss. For ResNet101 based model, masks selection can improve the segmentation performance obviously. Especially the best mIoU of 64.1 can be witnessed when $\tau = 0.8$ in SAL-Net-ResNet101. It is apparent that ResNet101 based network is less sensitive to the amount of annotations, but more sensitive to the quality of annotations. Then the refined mask scores SF defined in Eq. (7) are analysed and similar phenomena can be observed. As shown in Table IV, $\tau = 0.55$ can generate the best mIoU as 58.0 for SAL-Net-VGG16 and $\tau = 0.85$ can generate the best mIoU as 64.9 for SAL-Net-ResNet101. Therefore, $\tau = 0.55$ and $\tau = 0.85$ are used for defining the refined selection loss (defined in Eq. (13)) for VGG16 and ResNet101 based models respectively. It is noted that the refined masks scores SF perform better in selecting high-quality annotations than the saliency guided scores SC. For example, with the 3303 samples selected by SF, SAL-Net-ResNet101 can generate a mIoU of 64.9. However, with the 3638 samples selected by SC, SAL-Net-ResNet101 can only generate a mIoU 64.1. We can conclude that if a more powerful baseline model is selected, even a small amount of high-quality samples can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

TABLE III: Overall accuracy on PASCAL VOC 2012 test dataset. The IoUs of twenty categories and the mean IoU are presented.

Method	bk	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
SEC [51]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
Decouple-VGG16 [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56.4
BOOTSTRAPPING (VGG16) [14]	85.3	77.6	26.2	76.6	17.3	61.4	82.4	74.8	83.8	25.7	66.9	46.2	74.0	75.6	79.2	70.8	48.3	73.1	40.5	38.8	39.0	60.2
DSRG (VGG16) [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60.4
Pixel Affinity (VGG16) [13]	88.0	61.1	29.2	73.0	40.5	54.1	75.2	70.4	75.1	27.8	62.5	51.4	78.4	68.3	76.2	71.8	40.7	74.9	49.2	55.0	48.3	60.5
MDC (VGG 16) [30]	89.8	78.4	36.2	82.1	52.4	61.7	64.2	73.5	78.4	14.7	70.3	11.9	75.3	74.2	81.0	72.6	38.8	76.7	24.6	70.7	50.3	60.8
SeeNet (VGG16)[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60.7
FickleNet (VGG16) [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.9
SAL-Net-VGG16 (ours)	88.5	75.5	31.7	70.5	42.9	65.9	79.1	73.5	82.4	28	71.2	39.1	76.1	70.3	77.8	71	48	73.7	48.7	46.9	50.9	**62.5**
Decouple-ResNet101 [6]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60.1
Pixel Affinity (ResNet38) [13]	89.1	70.6	31.6	77.2	42.2	68.9	79.1	66.5	74.9	29.6	68.7	56.1	82.1	64.8	78.6	73.5	50.8	70.7	47.7	63.9	51.1	63.7
BOOTSTRAPPING (ResNet50) [14]	87.2	76.8	31.6	72.9	19.1	64.9	86.7	75.4	86.8	30.0	76.6	48.5	80.5	79.9	79.7	72.6	50.1	83.5	48.3	39.6	52.2	63.9
SeeNet (ResNet101)[7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62.8
DSRG (ResNet101) [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.2
FickleNet (ResNet101) [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.3
SAL-Net-ResNet101 (ours)	90.6	80	31.6	75.4	53.9	66.1	86.3	76.1	87.7	29.7	73.7	44.1	79.9	78.9	74.2	74.7	50.2	78.7	46.8	57.6	63	**66.6**

TABLE IV: Different values of thresholds τ are evaluated on PASCAL VOC 2012 validation dataset. The segmentation performance under different values of τ is reported for SAL-Net-ResNet101 and SAL-Net-VGG16. "NoS" represents the number of samples. The results of one single model are reported and the strategies such as multi-scale fusion and CRF are not used.

τ for SC						
Threshold-τ	**0.85**	**0.8**	**0.7**	**0.55**	**0.35**	**0**
NoS	2481	3638	5551	7535	9055	10490
SAL-ResNet101 mIoU	63.9	**64.1**	63.5	63.0	62.5	61.6
SAL-VGG16 mIoU	55.3	56.5	57.3	**57.6**	57.1	56.2
τ for SF						
Threshold-τ	**0.85**	**0.8**	**0.7**	**0.55**	**0.35**	**0**
NoS	3303	4362	5961	7586	8889	10392
SAL-ResNet101 mIoU	**64.9**	64.7	64.1	63.3	62.7	61.8
SAL-VGG16 mIoU	56.9	57.2	57.6	**58.0**	56.8	56.4

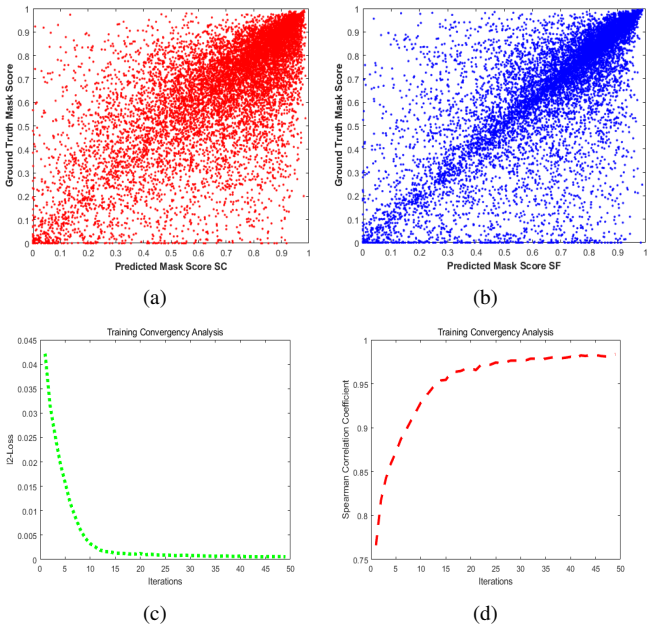

Fig. 9: Analysis on the performance of the mask scoring network. (a) Correlations between SC and their ground truth; (b) Correlations between SF and their ground truth; (c) The ℓ_2-loss in the training procedure; (d) The Spearman correlation coefficient in the training procedure.

generate satisfactory results. In order to evaluate the quality of network based mask scores SF more comprehensively, we plot each predictions and their corresponding ground truth in Fig. 9. The ground truth scores are obtained by calculating the AC (defined in Eq. (3)) using the ground truth pixel-level annotations. We can see that the refined mask scores SF have better correlation with their ground truth, especially for those prediction with values higher than 0.5. The correlation coefficients between the coarse mask scores SC and their ground truth is 70.1 for the initial coarse scores SC, while the correlation coefficient is 73.6 for network based refined scores SF. It indicates that the quality of the refined mask scores has been improved greatly via training the mask scoring network. Moreover, the ℓ_2-loss and the training accuracy defined by Spearman correlation coefficients of the mask scoring network in the training procedure are presented in Fig. 9 as well. We can see that the training procedure of mask scoring network almost converges around 30 epoches and a training accuracy around 0.984 can be achieved.

D. Analysis on the Classification Structures for Attention Loss

In this experiment, different network structures for the attention loss are evaluated by integrating various types of classification branches. As shown in Table V, the segmentation performance will be improved when multiple classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

TABLE V: Structure evaluations on the classification subnetwork for the attention loss. The results with multi-scale fusion and CRF are reported.

Settings	Rate	No Selection	Coarse Sel	Ref Sel
One branch	Rate=1	59.61	60.61	61.24
Two branch	Rate=1,3	59.71	60.65	61.26
Three branch	Rate=1,3,6	59.73	60.64	61.29
Four branch	Rate=1,3,6,9	59.75	60.69	61.33

TABLE VI: Ablation study on PASCAL VOC 2012 validation dataset and the mIoUs corresponding to P1 to P5 are presented. The values in "( )" indicates the results without multi-scale fusion and CRF. $\triangle$ denotes the cumulative improvements compared with P2.

SAL-Net-	**P1**	**P2**	**P3**	**P4**	**P5**
VGG16	56.9(52.1)	59.1(56.2)	60.2(57.6)	60.9(58.0)	61.3(58.4)
$\triangle_{vgg}$	-	-	1.1%(1.4%)	1.8%(1.8%)	2.2%(2.2%)
ResNet101	61.4(60.1)	62.2(61.5)	64.9(64.1)	65.8(64.9)	66.1(65.5)
$\triangle_{resnet}$	-	-	2.7%(2.6%)	3.6%(3.4%)	3.9%(4.0%)

branches are used, when the segmentation network is trained without selection loss. For different kind of classification network structures, training with coarse or refined selection loss jointly can both improve the segmentation performance. It is obvious that the architecture with four branches whose dilation rates are 1,3,6,9 can generate the best performance. When the four branch structure is used for constructing the attention loss, the proposed SAL-Net-VGG16 can achieve a mIoU of 61.3 when trained with the refined selection loss jointly.

E. Analysis on the Effects of Selection and Attention Losses for Ablation Study

The roles of five components are evaluated: P1: learning from image tags, P2: retraining with all the proxy annotations, P3: training with coarse selection loss, P4: training with refined selection loss and P5: +training with attention loss. Specifically, as shown in Table VI, the mIoUs of 56.9(52.1) and 61.4(60.1) are obtained for VGG16 and ResNet101 based models respectively in P1 by implementing DSRG [12]. Then the generated proxy annotations are applied to retrain the segmentation networks SAL-Net-VGG16 and SAL-Net-ResNet101 in P2. mIoUs of 59.1(56.2) and 62.2(61.5) can be achieved subsequently, and they are better than the reported 59.0 and 61.4 in [12]. In P3, the segmentation performance can be improved to 60.2(57.6) and 64.9(64.1) by selecting 7535 and 3638 high-quality annotations (Table IV) respectively. Therefore, the refined mask scores SF select 7586 and 3303 samples (Table IV) with higher quality to generate better segmentation performance as 60.9(58.0) and 65.8(64.9) respectively in P4. Finally, mIoU of 61.3(58.4) and 66.1(65.5) can be achieved by training the segmentation network with refined selection loss and attention jointly in P5. We can see that compared with the baseline P2, 1.1%(1.4%), 1.8%(1.8%) and 2.2%(2.2%) mIoU gains can be achieved for SAL-Net-VGG16 from P3 to P5 incrementally, and cumulative mIoU gains of 2.7% (2.6%), 3.6%(3.4%) and 3.9%(4.0%) can be

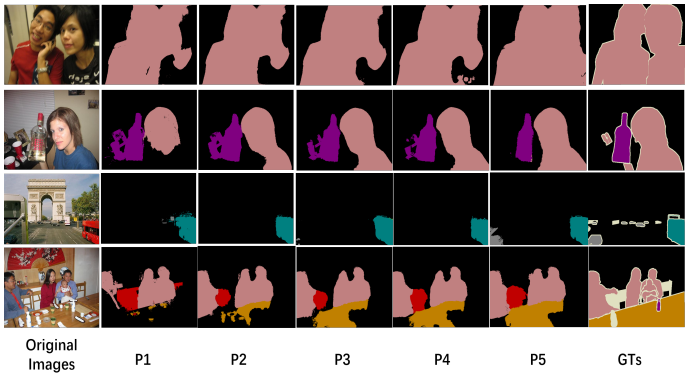

Fig. 10: Illustration of the segmentation results corresponding to ablation study steps from P1 to P5 in Table VI.

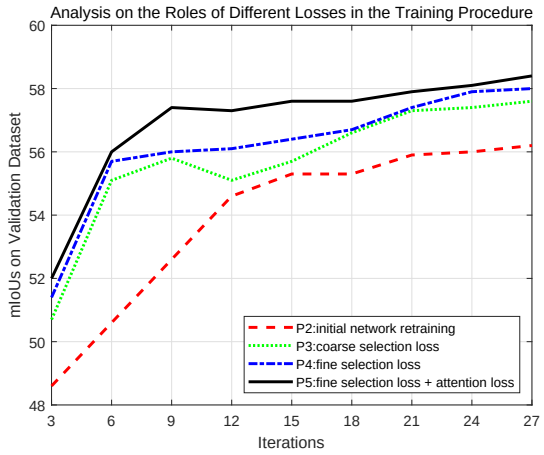

Fig. 11: Analysis on the roles of different losses for SAL-Net-VGG16 in the training procedure. The mIoUs on the validation set are reported every 3k iterations in the training process. The y-axis presents the mIoUs on the validation set and the x-axis presents the iterations from 3k to 27k.

obtained for SAL-Net-ResNet101 from P3 to P5. In summary, the segmentation performance with retraining is boosted from 59.1 to 61.3 and 62.2 to 66.1. Clearly, the significant performance improvements (2.2% and 3.9%) indicate that the proposed mechanisms of mask scoring and classification errors correction are effective in optimizing the segmentation networks, especially when the pixel-level annotations are noisy. Moreover, the qualitative results for ablation study are displayed in Fig. 10, which show that the segmentation masks corresponding to P1 to P5. It is observed that the segmentation masks are refined by discovering more discriminative regions and correcting classification errors adaptively. Then the effects of selection losses and attention loss are evaluated more comprehensively in Table VII. Without Sel loss and Atten loss, a mIoU of 59.1 is achieved by training from 10582 proxy annotations. The role of Sel loss is evaluated firstly, we can see that the coarse Sel loss can generate 1.1% performance gain and the refined selection loss can boost the performance to 60.9 with more reliable mask scores. When the Atten loss is integrated into the training framework, a mIoU of 59.8 is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2991592, IEEE Transactions on Multimedia

TABLE VII: Evaluating the roles of the selection loss and attention loss. The results with multi-scale fusion and CRF are reported.

Methods	**No Sel Loss**	**Coarse Sel Loss**	**Ref Sel Loss**
No Atten Loss	59.1	60.2	60.9
+Atten Loss	59.8	60.7	61.3

TABLE VIII: Segmentation performance on PASCAL VOC 2012 val dataset for different λ.

λ	**0**	**0.5**	**1**	**2**	**5**	**15**	**30**
$L_{sel} + \lambda L_{atten}$ mIoU	60.91	60.84	61.19	61.33	61.20	60.6	58.9

produced by taking the clean image tags as supervision even if the Sel loss is not integrated into the framework. Moreover, 0.9% and 1.5% mIoU gains can be witnessed when the whole framework is trained with coarse selection loss and refined selection loss respectively.

Furthermore, one additional experiment is performed to demonstrate the effectiveness of the proposed selection and attention losses. As shown in Fig. 11, the mIoUs on the validation set are recorded every 3k iterations in the training procedure. We can see that both the coarse and refined selection losses can improve the training efficiency, and a maximum advance mIoU as 1.9% can be recorded. When the selection loss is combined with the attention loss in P5, the training accuracy can climb to a high level within 6 epoches and reaches the highest value as 58.4 finally.

The improvement induced by the proposed losses is illustrated in Fig. 12, the SAL-Net with Sel loss and Atten loss can generate object masks of higher quality by adjusting the classification errors and refining the object boundaries adaptively. For example some pixels in the table region are miss-classified as bootle in the 1st image without the Atten loss, the the model trained with Atten loss can correct the wrongly labeled pixels by learning from the clean image tags. Furthermore, the segmentation model trained with Sel loss and Atten loss can generate segmentation masks with higher boundary accuracy by learning from more high quality supervisions, such as clean image tags or high-confident proxy annotations. In addition, we tried different values of λ in Eq. (14) to find the best performance for network training. The results for different values of λ are shown in Table VIII which show that setting $\lambda = 2$ in our method can achieve the best performance when simultaneously conducting mask selecting and segmentation probabilities refinement. We also found that the segmentation performance will drop sharply when λ is larger than 15. This is primarily because the attention loss will dominate the energy and the network tends to generate sparse localization maps so as to improving the classification accuracies, therefore, the segmentation performance will drop in consequence.

F. Experiments on Semi-supervised Semantic Segmentation

In order to evaluate the generality of the proposed selection loss and attention loss more comprehensively, the segmentation experiments with semi-supervised learning are conducted.

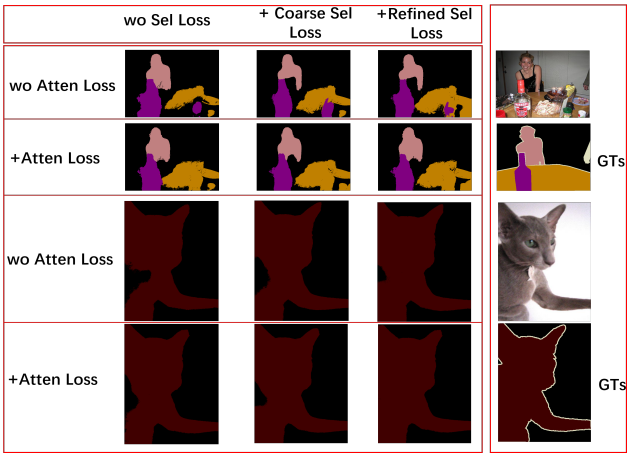

Fig. 12: Illustration of the roles of the attention loss and the selection loss. “wo” is the abbreviation of without. “Sel” is the short for selection. “Ref” is the short for refined. “Atten” is the short for attention.

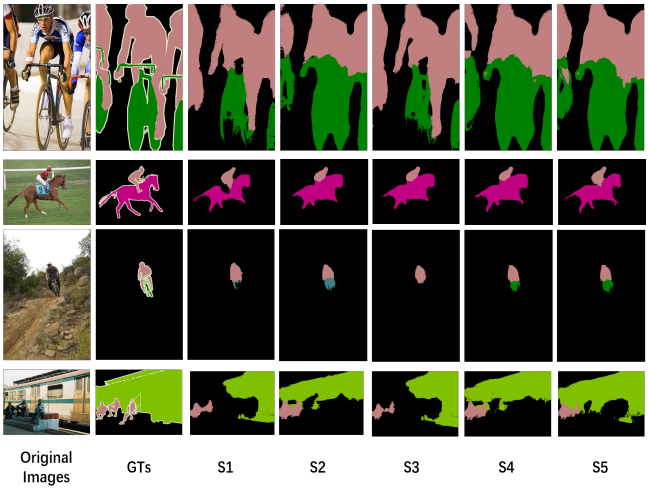

Fig. 13: Illustration of the effects of the selection and attention losses in semi-supervised semantic segmentation correspond to the results reported in Table IX.

In the semi-supervised setting, the segmentation models are trained on 1449 images with precise pixel-level annotations and other 9k images with image tags in the PASCAL VOC 2012 dataset. Different from the schemes proposed in DSRG [12] or FickleNet [8], SAL-Net-VGG16 utilizes the 9k initial proxy annotations and the image tags for network optimizing. In the first experiment, the effect of the selection and attention losses are evaluated, and the corresponding results are displayed in Table IX. In the setting of fully supervised segmentation with 1.4k images, the attention loss can boost the performance from 62.8 to 63.2. When the 9k proxy annotations are applied for learning, a mIoU of 64.5 can be achieved. Once the selection loss is utilized for annotations filtering, SAL-Net-VGG16 can obtain another 1.2% mIoU gain by selecting 7668 high quality samples. Finally, a mIoU of 66.2 is obtained by training the whole network with selection loss and attention loss jointly. The segmentation masks generated

TABLE IX: Evaluation of the selection and attention losses for semi-Supervised semantic segmentation.

S1	1.4kstrong	62.8
S2	1.4k strong+Atten loss	63.2
S3	1.4k strong+0.9k weak	64.5
S4	1.4k strong+0.9k weak+Sel Loss	65.7
S5	1.4k strong+0.9k weak+Sel Loss+Atten Loss	66.2

TABLE X: Comparison of semi-supervised semantic segmentation methods on PASCAL VOC 2012 validation set. The performances of DeepLab using 1.4K and 10.6K fully annotated data are presented as well.

Methods	Training Set	mIoU
DeepLab [47]	1.4K strong	62.5
WSSL [22]	1.4K strong + 9K weak	64.6
GAIN [52]	1.4K strong + 9K weak	60.5
MDC [30]	1.4K strong + 9K weak	65.7
DSRG [12] (baseline)	1.4K strong + 9K weak	64.3
FickleNet [8]	1.4K strong + 9K weak	65.8
SAL-Net-VGG16 (Proposed)	1.4K strong + 9K weak	**66.2**
DeepLab [47]	10.6K strong	67.6

by models trained with settings S1 to S5 are illustrated in Fig. 13. We can see that higher object boundary accuracies can be achieved when more high-quality annotations are generated and selected for training. SAL-Net-VGG16 is also compared with other state-of-the-arts such as FickleNet [8], DSRG [12], MDC [30] etc. The evaluation results shown in Table X have demonstrated that SAL-Net-VGG16 is more effective for semi-supervised semantic segmentation than other state-of-the-arts by generating high-quality supervision cues from image tags.

G. Analysis on the Computation Cost

As demonstrated in Table XI, the training & inference time of the segmentation networks and the mask scoring network are reported. Both networks are implemented by using Tensroflow on GTX 1080Ti. In the training stage, SAL-Net-VGG16 takes approximate 0.12s to process an image size 321×321 on average by employing selection loss and attention loss. Subsequently, the whole training procedure takes about 6.5 hours for 27000 iterations, while Deeplab-VGG16-ASSP [47] takes 6 hours. For SAL-Net-ResNet101, all the 15000 iterations during training period takes around 3.8 hours. As for the inference results, the inference of SAL-Net-VGG16 takes around $0.1s$ for segmenting an image with size 500×375 on average, while the inference of SAL-Net-ResNet101 consumes $0.15s$. Therefore, we found that the inference time of the proposed segmentation networks is the same as that of the baseline Deeplab-ASSP. Furthermore, the mask scoring network takes about $0.39s$ to process 8 inputs with size 432×432 in a batch. Afterwards, 516 seconds are consumed by an epoch and all 50 epoches (around 66k iterations) take 7.2 hours in total. The inference on an image with size 432×432 for SF of mask scoring network takes approximate 0.03s on average. Our proposed framework mainly focuses on the retraining procedure of weakly supervised segmentation. It consists of a mask scoring network and segmentation networks. It can be seen that the training & inference time of SAL-Net-VGG16 and SAL-Net-ResNet101 are almost the same as that of the baseline Deeplab-ASSP used in SEC and DSRG, but the segmentation performance has been significantly improved. The introduced extra computation costs are primarily caused by the training of mask scoring network, which is implemented in an offline manner. In summary, the complexity of designed algorithm has not been significantly increased when compared with the existing methods such as SEC and DSRG.

TABLE XI: The training & inference time of the mask scoring network and the segmentation networks.

	Bath Size	Iterations	Time Per Iteration	Training	Inference
Mask Scoring Network	8	66k	0.39s	7.2h	0.03s
SAL-Net-VGG16	7	27k	0.87s	6.5h	0.1s
SAL-Net-ResNet101	7	15k	0.91s	3.8h	0.15s
Deeplab-VGG16-ASSP	7	27k	0.81s	6h	0.1s

V. Conclusion

In this paper, we have introduced a novel end-to-end framework SAL-Net with two novel losses: a selection loss and an attention loss to effectively optimize the semantic segmentation model from weak and noisy annotations. Firstly, a coarse-to-fine mask scoring strategy is proposed to evaluate the quality of the produced segmentation masks in the training set. Then the selection loss is constructed to optimize the segmentation network on proxy annotations of high confidence. Moreover, an attention loss is proposed to learn the attention weights of each class from clean image tags. The classification ambiguities are refined adaptively by reciprocally learning from clean image tags and interacting with the segmentation network. The proposed framework SAL-Net is evaluated on the challenging PASCAL VOC 2012 benchmark, extensive numerical and visualization results demonstrate the benefits brought by the proposed new losses. In the future, we will explore enhancing the performance of SAL-Net for images with more complex scenes.

References

[1] Yi Li, Yanqing Guo, Jun Guo, Zhuang Ma, Xiangwei Kong, and Qian Liu, "Joint crf and locality-consistent dictionary learning for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 875–886, 2018.

[2] Hengcan Shi, Hongliang Li, Fanman Meng, Qingbo Wu, Linfeng Xu, and King Ngi Ngan, "Hierarchical parsing net: Semantic scene parsing from global scene to objects," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2670–2682, 2018.

[3] Byeongkeun Kang, Yeejin Lee, and Truong Q Nguyen, "Depth-adaptive deep neural network for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2478–2490, 2018.

[4] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille, "Pcl: Proposal cluster learning for weakly supervised object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 176–191, 2018.

[5] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[6] Tianyi Zhang, Guosheng Lin, Jianfei Cai, Tong Shen, Chunhua Shen, and Alex C Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2930–2941, 2019.

[7] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng, "Self-erasing network for integral object attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 547–557.

[8] 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 conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 5267–5276.

[9] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7223–7233.

[10] Alexander Kolesnikov and Christoph H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 695–711.

[11] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," *arXiv preprint arXiv:1707.05821*, 2017.

[12] Zilong Huang, Xinggang Wang, Jiashi 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 (CVPR)*, 2018, pp. 7014–7023.

[13] Jiwoon Ahn and Suha Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7014–7023.

[14] Tong Shen, Guosheng Lin, Chunhua Shen, and Ian Reid, "Bootstrapping the performance of webly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1345–1362.

[15] 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 (CVPR)*, 2018, pp. 1325–1334.

[16] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang, "Self-produced guidance for weakly-supervised object localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 597–613.

[17] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1345–1362.

[18] 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 International Conference on Computer Vision (ICCV)*, 2019, pp. 2070–2079.

[19] B Jin, M.V.O Segovia, and S Süsstrunk, "Webly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1705–1714.

[20] S Hong, D Yeo, S Kwak, H Lee, and B Han, "Weakly supervised semantic segmentation using web-crawled videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2224–2232.

[21] Wataru Shimoda and Keiji Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5208–5217.

[22] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1742–1750.

[23] 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 (CVPR)*, 2016, pp. 3159–3167.

[24] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4410–4419.

[25] 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 (CVPR)*, 2016, pp. 2921–2929.

[26] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," *arXiv preprint arXiv:1703.08448*, 2017.

[27] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon, "Two-phase learning for weakly supervised object localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3534–3543.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[29] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[30] 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 (CVPR)*, 2018, pp. 7268–7277.

[31] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[32] Zhiwu Lu, Zhenyong Fu, Tao Xiang, Peng Han, Liwei Wang, and Xin Gao, "Learning from weak and noisy labels for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 486–500, 2016.

[33] Aoxue Li, Zhiwu Lu, Liwei Wang, Peng Han, and Ji-Rong Wen, "Large-scale sparse learning from noisy tags for semantic segmentation," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 253–263, 2016.

[34] Suha Kwak, Seunghoon Hong, and Bohyung Han, "Weakly supervised semantic segmentation using superpixel pooling network.," in *AAAI*, 2017, pp. 4111–4117.

[35] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," *arXiv preprint arXiv:1807.11719*, 2018.

[36] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2314–2320, 2017.

[38] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3203–3212.

[39] Yutaka Sasaki et al., "The truth of the f-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[40] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3286–3293.

[41] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 784–799.

[42] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6409–6418.

[43] K. M. He, X. Y. Zhang, S.Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[44] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "Density modeling of images using a generalized normalization transformation," *arXiv preprint arXiv:1511.06281*, 2015.

[45] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[46] Debidatta Dwibedi, Ishan Misra, and Martial Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1301–1310.

[47] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[48] 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*, vol. 88, no. 2, pp. 303–338, 2010.

[49] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Semantic contours from inverse detectors," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 991–998.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] A Kolesnikov and C.H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 695–711.

[52] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu, "Tell me where to look: Guided attention inference network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9215–9223.

Lei Zhou received the B.S. degree from Wuhan University of Technology, Wuhan, China, in 2008, and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China in 2014, under the supervision of Prof. Jie Yang. He is currently a lecturer with School of Medical Instrument and Food Engineering, and is also with Shanghai Engineering Research Center of Assistive Devices, University of Shanghai for Science and Technology, Shanghai, China. His research group has won 6 championships in CVPR 2018 and CVPR 2019 CLIC image compression challenges. His current research interests include semantic segmentation, medical image analysis and image/video compression.

Chen Gong received his B.E. degree from East China University of Science and Technology (ECUST) in 2010, and dual doctoral degree from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS) in 2016 and 2017, respectively. Currently, he is a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly include machine learning, data mining, and learning-based vision problems. He has published more than 80 technical papers at prominent journals and conferences such as IEEE T-PAMI, IEEE T-NNLS, IEEE T-IP, IEEE T-CYB, IEEE T-CSVT, IEEE T-MM, IEEE T-ITS, ACM T-IST, NeurIPS, CVPR, AAAI, IJCAI, ICDM, etc. He also serves as the reviewer for more than 20 international journals such as AIJ, IEEE T-PAMI, IEEE T-NNLS, IEEE T-IP, and also the SPC/PC member of several top-tier conferences such as ICML, NeurIPS, CVPR, AAAI, IJCAI, ICDM, AISTATS, etc. He received the "Excellent Doctorial Dissertation" awarded by Shanghai Jiao Tong University (SJTU) and Chinese Association for Artificial Intelligence (CAAI). He was enrolled by the "Young Elite Scientists Sponsorship Program" of Jiangsu Province and China Association for Science and Technology. He was also the recipient of "Wu Wen-Jun AI Excellent Youth Scholar Award".

Zhi Liu (M'07-SM'15) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002 and 2005, respectively. He is currently a Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 190 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC member/session chair in ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, WIAMIS 2013, etc. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of Signal Processing: Image Communication and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in Signal Processing: Image Communication.

Keren Fu received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2011, and dual Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, and Chalmers University of Technology, Gothenburg, Sweden, in 2016, under the joint supervision of Prof. Jie Yang and Prof. Irene Yu-Hua Gu. He is currently a research associated professor with College of Computer Science, Sichuan University, Chengdu, China. His current research interests include visual computing, saliency analysis, and machine learning.