Scribble-Supervised Semantic Segmentation Inference

Jingshan Xu[1*], Chuanwei Zhou[1*], Zhen Cui[1†], Chunyan Xu[1†], Yuge Huang[2], Pengcheng Shen[2], Shaoxin Li[2], Jian Yang[1]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2]Tencent, Shenzhen, China

Abstract

In this paper, we propose a progressive segmentation inference (PSI) framework to tackle with scribble-supervised semantic segmentation. In virtue of latent contextual dependency, we encapsulate two crucial cues, contextual pattern propagation and semantic label diffusion, to enhance and refine pixel-level segmentation results from partially known seeds. In contextual pattern propagation, different-granular contextual patterns are correlated and leveraged to properly diffuse pattern information based on graphical model, so as to increase the inference confidence of pixel label prediction. Further, depending on high-confidence scores of estimated pixels, the initial annotated seeds are progressively spread over the image through dynamically learning an adaptive decision strategy. The two cues are finally modularized to form a close-looping update process during pixel-wise label inference. Extensive experiments demonstrate that our proposed progressive segmentation inference can benefit from the combination of spatial and semantic context cues, and meantime achieve the state-of-the-art performance on two public scribble segmentation datasets.

1. Introduction

Semantic segmentation is one fundamental topic in computer vision. Numerous deep learning based methods have sprung up to deal with this task [5, 6, 28, 10, 43]. They usually require a vast quantity of fully annotated samples to facilitate the training of deep neural networks. But the annotation of image segmentation often suffers massive workload because of fully flexible/irregular polygons of segmentation regions. To bypass the dependency on high-expensive annotations, weakly supervised semantic segmentation is much desirable due to more convenient annotations. Broadly speaking, the annotation ways mainly contain four categories: image-level labels [9, 18, 24, 41], clicks [4], bounding boxes [7, 30, 19] and scribbles [26, 34, 35, 38]. The image-level labels and clicks only provide very limited supervision information, thus making them hard to train high-accuracy semantic segmentation models. Although bounding boxes can provide more supervision information, they tend to overlap with each other and thus result into many disturbances from confusion labels during training. In contrast, the scribble annotations are more flexible to reflect the distribution of semantic classes, and their operation is easily controllable. Due to its potential advantages, here we focus on the case of scribble-supervised semantic segmentation.

The scribble-supervised semantic segmentation has been studied over the past decade. Early methods [33, 12] may be dated to interactive segmentation which uses graphical models to directly expand semantic labels to unlabeled regions. With the popularity of deep neural networks, many recent methods attempted to introduce deep feature learning into traditional graphical models, e.g., CRF [26] and random-walks [37] based on scribbles [26], and generated more confident pseudo labels for those unlabeled regions to guide model update. Further, due to the difficulty of boundary estimation, some methods [38] employed auxiliary networks to aid segmentation refinement, or introduced topology-constrained loss functions [34, 35] to smooth prediction results. Although these methods endeavored to utilize more robust features or external information to improve segmentation performance, the crucial problems about what/how to infer from known seeds to unknown regions especially in the deep feature pattern space are still under studied.

To address the above problems, we propose a progressive segmentation inference (PSI) framework by adaptively diffusing contextual patterns as well as label information for scribble-supervised semantic segmentation. Inspired by the observation [27] that the patterns from low-level visions to high-level semantics are mutually dependent/correlated in spatial domain or semantic domain, we attempt to leverage the pattern dependency to fulfill segmentation infer-

*Authors contributed equally.
†Corresponding authors: zhen.cui@njust.edu.cn, cyx@njust.edu.cn.

ence. Concretely, the designed segmentation inference consists of two components: multi-granular contextual pattern propagation (CPP) and progressive semantic label diffusion (SLD). In CPP, pattern correlations are mined from multi-granular contextual domain, including across different convolutional layers as well as across different spatial positions. Afterwards, the pixel-level feature patterns to predict pixel labels are enhanced through the information aggregation in the multi-granular contextual domain through graphical model. Attributed to the introduction of graph structure, the advantages of CPP are two folds. On the one hand, the contextual information from different-granular layers could be effectively integrated to better estimate pixel labels. On the other hand, during training the backward gradients could be propagated to former image pixels more quickly due to shorted connections in graph, which also makes scarce supervision of top layers more effectively imposed on bottom layers. In SLD, motivated by [17, 42], we attempt to extend the annotated scribbles to neighboring confident areas according to the prediction scores of the segmentation model with CPP. To make sure more confident inference, we introduce an adaptive decision strategy to choose those high-confidence regions as pseudo ground truths, which are further used for the next model update process. Hereby the unknown segmentation labels could be gradually refined with continuous extensions of high-confidence regions. The two components are modularized and further encapsulated into a close-looping inference process to fulfill progressive segmentation prediction. The extensive experiments show that our proposed PSI method could mutually evolve segmentation model as well as labels, and meantime achieve the state-of-the-art results in the task of scribble-supervised semantic segmentation.

In summary, our contributions are three folds:

- We propose a novel progressive segmentation inference framework through context inference as well as annotation inference for scribble-supervised semantic segmentation.
- We develop two crucial components, multi-granular contextual pattern propagation and progressive semantic label diffusion, to form a close-looping update progress during pixel-label inference.
- We experimentally validate the effectiveness of the proposed two components, and report the state-of-the-art performance.

2. Related Works

Semantic Segmentation. In the early years, the semantic segmentation methods mainly employed the graphical models such as CRF [22]. The literature [11] integrated a conditional graphical model and location priors to produce semantic segmentation results. [21] introduced the dense-CRF and made it computationally efficient through the permutohedral lattice [1]. With the rapid development of deep learning, many deep segmentation networks have been developed. FCN [28] first introduced the deep networks into semantic segmentation and achieved a great performance boost. SegNet [3] developed an encoder-decoder structure with deconvolution as well as unpooling layers and dropped the fully connected layer utilized in FCN [28]. Afterwards, the contextual information excavation aroused great attentions and many methods were developed to utilize the contextual information priors. DeepLab [5] employed the atrous convolution to enlarge the receptive field of the convolution kernels to perceive a broader contextual region and used CRF [21] to refine the segmentation predictions. CENet [45] exploited an end-to-end trainable neural network to learn context encoding vectors. DeepLabV3+ [6] aggregated the ASPP module and encoder-decoder structure to further enlarge the receptive field. DANet [10] utilized the self-attention [36] to aggregate the global contexts both from the spatial and the channel dimension. OCR [43] employed HRNet [39] as the backbone and utilized a coarse segmentation result to obtain the object regional contexts to further refine the segmentation result. A majority of the semantic segmentation researches attempt to employ more contextual information, so as ours. Different from the previous methods, our method developes an effective mechanism for the aggregation of contexts both in spatial and semantic domains through graphical model.

Scribble-Supervised Segmentation. In the early stage, scribble-supervised segmentation was usually addressed in an interactive manner [33, 12] where feedback scribbles were continuously drawn for refining the segmentation results. Methods in this stage usually converted an image to a weighted undirected graph. With the surge of deep learning, many researches have attempted to utilize deep neural networks to address the scribble-supervised segmentation. ScribbleSup [26] first introduced deep learning into the scribble-supervised segmentation. A full annotation map was first generated using the weakly annotated scribbles and a CRF model [21]. Afterwards, the optimization of the neural network and the CRF energy function were alternately implemented to refine the segmentation results. RAWKS [37] embedded a deep segmentation network and a learnable label-propagator to progressively update the segmentation network and the propagated dense annotations. Our proposed PSI also jointly updates the segmentation network and the annotation maps, but our annotations are gradually expanded to the unlabeled regions according to a dynamically learned strategy. BPG [38] developed a perception refinement network to utilize more information from the encoder especially from the larger resolution feature maps. BPG also specifically designed an auxiliary network to refine the edge details which was trained under extra

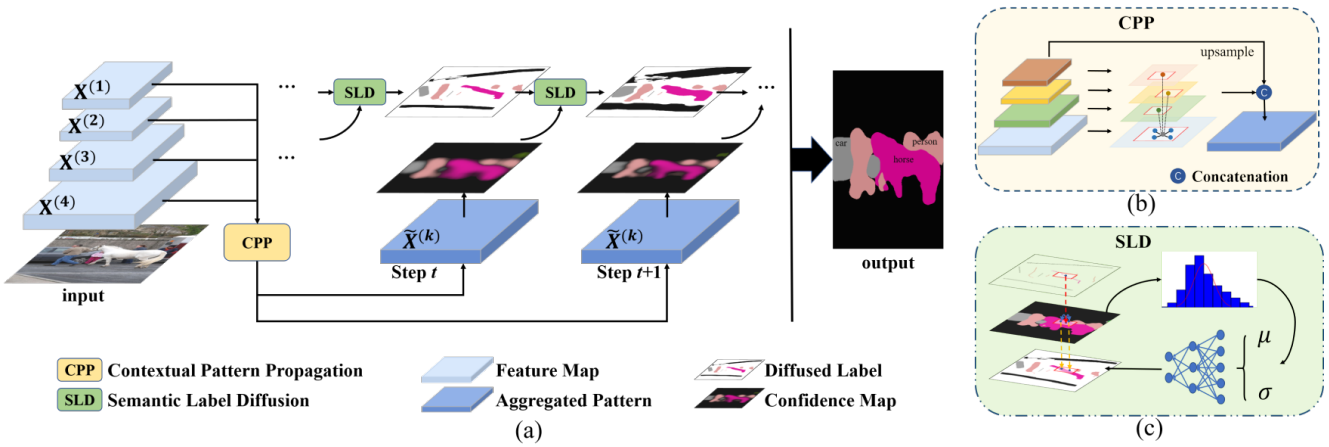

Figure 1. Illustration of our proposed progressive segmentation inference method. (a) The overall framework. (b) A detailed example of our proposed CPP module. (c) Detailed structure of our proposed SLD module. Our method mainly consists of two components: the contextual pattern propagation (CPP) module and the semantic label diffusion (SLD) module. In CPP, multi-scale contextual patterns are propagated through graphical model both in spatial and semantic domains to form the enhanced aggregation pattern. In SLD, a dynamic decision network is designed to adaptively diffuse semantic labels to unlabeled regions according to distributions of the estimated confidence maps. CPP and SLD are finally modularized into a close-looping to progressively update the segmentation network and the supervisions.

edge information. Our proposed PSI also attempts to capture more internal information like [47] but we aggregate different-granular contextual patterns in spatial and semantic domains, and we don't resolve to either larger resolutions nor other extra information. Other researches [34, 35] focused on designing topology-constrained loss functions to constrain the network learning with only scribble annotations. In contrast, our proposed PSI mainly devotes to design novel network structures to perform effective inference of the segmentation results as well as the annotations, and we only use the simple cross entropy and $\mathcal{L}_1$ penalties to train our networks.

3. Progressive Segmentation Inference

3.1. Overview

The framework of our proposed progressive segmentation inference is depicted in Fig. 1. Given a set of training samples with weakly-annotated scribbles, our aim is to learn a more robust segmentation model. Due to the limited label information of scribbles, there usually lacks sufficient guidance to train an excellent segmentation model. To circumvent this problem, on the one hand, internal/self-priors of images (such as contextual pattern correlations or topological structure information) need to be fully excavated for learning. On the other hand, label information is supposed to be inferred and diffused from the annotated scribbles to unlabeled pixels, so as to furnish serviceable annotations as much as possible. To this end, we encapsulate two crucial cues, contextual pattern propagation (CPP) and semantic label diffusion (SLD), to enhance and refine pixel-level segmentation results from weakly-annotated scribbles.

Given an input image I, we suppose the scribble annotation region as $\mathcal{R}_0$, and the corresponding labels as $y_{\mathcal{R}_0}$ aka initial seeds. We attempt to infer confident supervision labels from the initial region $\mathcal{R}_0$ to the whole image $\mathcal{R}_I$ by mining internal contextual cues from I. First we encode the input image I with some popular convolutional neural networks (CNN) such as ResNet [15], and generate multi-scale feature maps $\{\mathbf{X}^{(i)}\}|_{i=1}^{L}$, where the superscript (i) denotes the layer index, $\mathbf{X}^{(i)}$ denotes higher-level semantic feature map with a bigger i. The segmentation region $\mathcal{R}_t$ (at the initial $t = 0$) would be expanded with the progressive evolution processes. In each stage, we leverage CPP to deeply exploit the internal contextual pattern relevance of the input image I based on graph topological structure, and aggregate the contextual patterns in semantic and spatial domains. Suppose the expected target size of the enhanced feature map is equal to the scale of the k-layer's feature map and call it the k-th destination layer. The enhanced feature map can be derived as $\widetilde{\mathbf{X}}^{(k)} \leftarrow \text{CPP}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots)$. The details of CPP will be introduced in Section 3.2. Once the enhanced contextual pattern $\widetilde{\mathbf{X}}^{(k)}$ obtained, we can predict a confidence map M through one regression function (e.g. a convolutional network layer). The confidence map M together with the previous seed $y_{\mathcal{R}_t}$ are fed into the SLD stage to infer the new seed region denoted as $y_{\mathcal{R}_{t+1}}$. To better predict the new seeds, we specifically design a dynamic strategy network to adaptively expand the supervision areas according to the distribution of confidence scores in the estimated segmentation maps. More details about SLD could be found in Section 3.3. Therefore, the new seed $y_{\mathcal{R}_{t+1}}$ may be used as the supervision information to guide the next training process and further continue to grow the seeds.

CPP and SLD are finally modularized and further encapsulated into a close-looping inference process to optimize the segmentation network as well as segmentation labels.

3.2. Contextual Pattern Propagation

The goal of the CPP module is to further exploit internal contextual pattern relevance of the input image both in semantic and spatial domains by modeling the topological structure of patterns with graph, enhancing pattern representations with image's self-priors. In semantic domain, the contextual patterns from low-level detailed feature layers to high-level semantic feature layers are captured; In spatial domain, the contextual patterns in the certain layer are aggregated locally. Specifically, we model the CPP aggregation process by a graph $G = (\mathcal{V}, \mathcal{A}, \mathcal{X})$ defined on multi-granularity feature layers, where $\mathcal{V}, \mathcal{A}, \mathcal{X}$ are the node set, adjacency matrix and node attributes/features respectively. For simplification, below we take a local subgraph around one node to illustrate the CPP module.

Suppose we want to derive the pattern at one spatial position of the destination layer k, where this spatial position corresponds to one node denoted as $v_i^{(k)} \in \mathcal{V}$. Given the destination node $v_i^{(k)}$, we can construct multi-granular adjacent relations across different feature levels. The neighbor node set w.r.t $v_i^{(k)}$ is denoted as $\mathcal{N}_i^{(k)}$, which also contains the self-node $v_i^{(k)}$. According to the neighbor node set, we can construct the corresponding adjacency matrix $\mathbf{A}_i^{(k)}$ and the attribute matrix $\mathbf{X}_i^{(k)}$. The pattern aggregation can be formulated as:

$$\widetilde{\mathbf{x}}_i^{(k)} = f_{\text{agg}}(\mathbf{X}_i^{(k)}, \mathbf{A}_i^{(k)}, \Theta_p), \tag{1}$$

where $\widetilde{\mathbf{x}}_i^{(k)}$ is the enhanced patterns w.r.t the destination node, Θ_p is the model parameter to be learnt. Next, we introduce the construction of the adjacency matrix $\mathbf{A}_i^{(k)}$ and the aggregation function f_{agg}.

To determine the neighbor nodes w.r.t the destination node, we define an operator $p^{(l)}(\cdot)$ as:

$$p^{(l)}(i) = \begin{cases} \lfloor \dfrac{i}{2^{S_k - S_l}} \rfloor, & \text{if} \quad S_k > S_l, \\ i \cdot 2^{S_l - S_k}, & \text{otherwise.} \end{cases} \tag{2}$$

where the operator $\lfloor \cdot \rfloor$ means rounding down. S_k is the scale of the destination layer $\mathbf{X}^{(k)}$, S_l is the scale of the aggregation layer $\mathbf{X}^{(l)}$. For instance, given $\mathbf{X}^{(k)} \in \mathbb{R}^{C_k \times H_k \times W_k}$ and $\mathbf{X}^{(l)} \in \mathbb{R}^{C_l \times 2H_k \times 2W_k}$, then the scale of the destination layer k is considered as 1 and the scale of the aggregation layer l is 2. We then define the adjacency nodes set $\mathcal{N}_i^{(k)}$ of $v_i^{(k)}$ over all correlation layers:

$$\mathcal{N}_i^{(k)} = \{ Idx^{(l)}(j) \mid |p^{(l)}(i) - Idx^{(l)}(j)| < \rho_i^{(l)}, l \in \mathcal{I} \}, \tag{3}$$

where $Idx^{(l)}(j)$ denotes the specific position of an adjacency node $v_j^{(l)}$ in the correlation layer $\mathbf{X}^{(l)}$, $\rho_i^{(l)}$ denotes the radius of the contextual neighbor window in the l-th layer, $\mathcal{I}$ is the set of all correlation layers indexes.

According to the neighbor node set $\mathcal{N}_i^{(k)}$, we can construct the adjacency matrix $\mathbf{A}_i^{(k)}$. To calculate the edge weights between nodes, we introduce a relationship metric $\langle \cdot \rangle$ and it could be Euclidean, cosine distances or the inner product. In this paper, we utilize the inner product as our relationship metric, so the edge weight $A_{i,j}^{(k)}$ of the destination node $v_i^{(k)}$ and the adjacency node $v_j^{(l)}$ is equal to $\langle \mathbf{x}_i^{(k)}, \mathbf{x}_j^{(l)} \rangle$, then we can obtain the normalized edge weight $A_{i,j}^{(k)} \leftarrow \frac{\exp\{A_{i,j}^{(k)}\}}{\sum_{m \in \mathcal{N}_i^{(k)}} \exp\{A_{i,m}^{(k)}\}}$. When $k \neq l$, the contextual patterns are aggregated in different-layer semantic domain, low-level detaileds and high-level semantics are fused and complementary from different-granularity; When $k = l$, the contextual patterns are aggregated in spatial domain, and spatial patterns in the same semantic layer are propagated to enhance the node representations.

Going back to our aggregated pattern of the node $v_i^{(k)}$, we can now reformulate the Eqn. (1) as:

$$\widetilde{\mathbf{x}}_i^{(k)} = \sigma\Big(\sum_{l \in \mathcal{I}} \sum_{v_j \in \mathcal{N}_i^{(k,l)}} A_{i,j}^{(k)} \cdot \mathbf{x}_j^{(l)} \cdot \mathbf{W}^{(l)} + \mathbf{b}^{(l)}\Big), \tag{4}$$

where $\mathcal{N}_i^{(k,l)} \subseteq \mathcal{N}_i^{(k)}$ is the neighbor node set at the l-th layer, the parameters $\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}$ are used to aggregate contextual patterns of different-granularity feature maps into a uniform semantic domain, and σ is one nonlinear activation function.

Following DeepLabV3+ [6], we bilinearly upsample the aggregation pattern $\widetilde{\mathbf{x}}_i^{(k)}$ and the highest-level feature map to the low-level feature map's size and then concatenate them with the corresponding low-level patterns, forming an effective decoder module. It is clear that the advantage of our proposed CPP module are two folds. On the one hand, the multi-scale contextual patterns of various granularity could be effectively aggregated in assistance of graphical model to enhance pattern representations, promoting the semantic labels inference. On the other hand, during training the backward gradients could be propagated to former image pixels more quickly due to shorted connections in graph, which also makes scarce supervision of top layers more effectively imposed on bottom layers. Our experiment results have shown that our proposed CPP module is able to outperform a strong baseline and leads to satisfactory semantic segmentation results.

3.3. Semantic Label Diffusion

After aggregating contextual patterns in spatial and semantic domains, we can predict a confidence map M through a segmentation network $F_{\text{seg}}(I, \Theta)$ with learnable

parameters $\Theta = \Theta_p \cup \Theta_s$, where Θ_s is the parameter of segmentation model. Although scribbles provide annotations to some extent, the learned segmentation network's performance is still limited due to the scarcity of supervisions. In effect, due to the intrinsic spatial consistency of natural images, the weak supervisions can be expanded properly according to the estimated confidence scores. To this end, we propose a progressive semantic label diffusion process through a differentiable dynamic decision network $F_{\text{sld}}(\cdot)$. In the step t, diffused labels $y_{\mathcal{R}_{t+1}}$ are obtained:

$$y_{\mathcal{R}_{t+1}} = F_{\text{sld}}(y_{\mathcal{R}_t}, M_t, \theta) = \cup(y_{\mathcal{R}_t}, \mathcal{B}(M_t, \theta)), \tag{5}$$

where $y_{\mathcal{R}_t}$ is the supervision map at the t-th step and when $t = 0$ it represents for the original seeds. $\mathcal{B}(\cdot)$ indicates a binarization process with learnable parameters θ, $\cup$ represents the union operator. As is indicated, weak annotations could be further enhanced through the union of the previous supervision $y_{\mathcal{R}_t}$ and the binarized confidence map M_t, leading to the diffused label supervision $y_{\mathcal{R}_{t+1}}$ which could be further utilized for the next optimization step of the segmentation network.

For a conventional method, the expansion principle is usually heuristic which manually sets a hard threshold γ to perform the binarization process $\mathcal{B}(\cdot)$. However, the manual threshold setting requires many validation trials and the hard binarization is not differentiable. Even worse, the hard binarization ignores the intrinsic characteristics of the images. In effect, the binarization process should be adaptively determined according to the property of the estimated confidence map M. To this end, we design a strategy network to dynamically generate a soft binarization map based on the distribution of confidence scores. Supposing at the t-th step, for each semantic class c, the supervision region is expanded based on the current supervision region as $\Omega_{c,r}$ with a growing window of radius r, the mean μ_c and variance σ_r^2 of the confidence distribution are first computed by:

$$\mu_c = \frac{1}{|\Omega_{c,r}|} \sum_{i \in \Omega_{c,r}} M(c, i), \tag{6}$$

$$\sigma_c^2 = \frac{1}{|\Omega_{c,r}|} \sum_{i \in \Omega_{c,r}} (M(c, i) - \mu_c)^2, \tag{7}$$

where $M(c, i)$ means the confidence score of the class c at the position i. The computed mean μ_c and variance σ_c^2 are utilized to generate a dynamic threshold γ_c. Note that the growing window radius r increases as the step t. For the presentation convenience, we drop the subscript c in the following contents. We then design a small network with learnable parameter θ to adaptively learn a dynamic threshold $\gamma(\theta)$ by inputing the computed μ and σ:

$$\gamma(\theta) = \mu + g(\mu, \sigma^2|\theta) \cdot \sigma, \tag{8}$$

where the operator g means the forward process of the network θ.

The hard binarization is still an obstacle to keep the network θ from learnable. We then introduce a soft binarization technique and achieve a differentiable adaptive decision strategy:

$$\mathcal{B}(M_t, \theta) = \frac{1}{1 + \exp\{-k \cdot \max(M_t - \gamma(\theta), 0)\}}, \tag{9}$$

where k is a factor which we set to be 20. Another issue comes from the union operator $\cup$ in the Eqn. 5 and here we employ a max operator to make it differentiable.

With one step semantic label diffusion, the updated supervisions could provide more accurate and adequate guidance for the learning of the segmentation network, then the confidence maps predicted in the next iteration will be promoted which will further result in a new round semantic label diffusion. Therefore, the proposed segmentation network with CPP and label diffusion network with SLD could be encapsulated into a close-looping to mutually evolve the segmentation network parameters and the supervisions. Our experimental results have shown that our proposed semantic label diffusion module can progressively update the supervision maps and lead to better segmentation results, besides, it's compatible with the proposed CPP module and the combination of the two modules could further lead to more promoted results.

4. Optimization Objective

We design two different objective functions $\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{sld}}$ for optimizations of the segmentation network F_{seg} with Θ and the label diffusion network F_{sld} with θ, respectively. The pixel-wise cross entropy is mainly employed as one part of the loss functions. The cross entropy is imposed over the labeled regions $\mathcal{R}$, formally:

$$\mathcal{L}_{\text{ce}}(p, y) = -\frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \sum_{c=1}^{C} y_{i,c} \log p_{i,c}, \tag{10}$$

where $y_{i,c}$ denotes the ground-truth in position i for each semantic category c, $p_{i,c}$ donates the predicted semantic scores in position i for the class c and C is the total number of the semantic categories.

For the segmentation network optimization, the update of the parameters Θ may suffer from the sparsity of the scribble annotations especially at the early stages. To mitigate the above issue, we further introduce a smoothness penalty to constrain the network learning. For each location, we adopt the mean $\mathcal{L}_1$ distance between the prediction and its 8-neighbors. Define the smooth penalty term averaged over all positions as $\mathcal{L}_p$, our segmentation loss function $\mathcal{L}_{\text{seg}}$ at t-th step is defined:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{ce}}(M_t, y_{\mathcal{R}_t}) + \lambda_1 \mathcal{L}_p(y_{\text{seg}}), \tag{11}$$

where M_t is the normalized semantic scores predicted by the segmentation network, $y_{\mathcal{R}_t}$ is the ground-truth given at t-th step, $y_{\text{seg}} = \arg\max(M_t)$ is the predicted label map, and λ_1 is a balance factor.

For the label diffusion network optimization, the update of the parameters θ of SLD are constrained by two terms:

$$\mathcal{L}_{\text{sld}} = \mathcal{L}_{\text{ce}}(y_{\mathcal{B}_t}, y_{\mathcal{R}_0}) + \lambda_2 \mathcal{L}_{\text{ce}}(y_{\text{seg}}, y_{\mathcal{B}_t}), \qquad (12)$$

where $y_{\mathcal{B}_t} = \mathcal{B}(M_t, \theta)$ is the soft label map predicted by SLD, $y_{\mathcal{R}_0}$ is the initial scribble labels and $\lambda_2 = 1$ is a balance factor.

We present an alternating solution to optimize the segmentation network with CPP and the label diffusion network with SLD. At t-th optimization step, we first fix θ and solve for Θ. Through the segmentation network, we can obtain a confidence score map M_t and a coarse segmentation mask y_{seg}, then Θ is updated by $\mathcal{L}_{\text{seg}}$ with the given ground-truth $y_{\mathcal{R}_t}$. Next, we fix Θ and solve for θ. With the output from the segmentation network, the soft binarization map $y_{\mathcal{B}}$ can be estimated by SLD, then $\mathcal{L}_{\text{sld}}$ is used to update the label diffusion network. Eventually, the two networks are modularized to form a close-looping update process, achieving sufficient pattern excavation and label inference.

5. Experiment

5.1. Experimental setup

The proposed method is evaluated on the PASCAL VOC 2012 Semantic segmentation dataset [8] and the PASCAL Context dataset [29]. The PASCAL VOC 2012 dataset contains in total 20 foreground classes and one background class. We train our network on the extended training set following [13] which includes 10582 training images and evaluate the performance on the validation set which contains 1449 fully annotated images. Additionally, we conduct experiments on the PASCAL Context dataset (4998 training images, as well as 5105 validation images with 59 classes and one background class) to verify the performance of our approach. All scribble supervisions during training are from [26]. We adopt the mean Intersection-over-Union (mIoU) score as our evaluation metric. All experiments are implemented in the Pytorch Framework [31]. Unless otherwise specified, all ablation studies are conducted on the PASCAL VOC 2012 dataset [8].

5.2. Implementation details

We adopt the ResNet101-based DeepLabV3+[6] as the backbone and the output stride is 16. We re-train deeplabV3+ with scribble supervisions and take it as our baseline. We take the output feature maps generated by the Conv2 - Conv5 Blocks as the aggregation maps and we set $\mathbf{X}^{(4)}$ as the destination scale map. Each feature map is processed by a 1×1 convolution before CPP to uniform the semantic domain. The contextual neighbor radiuses ρ^l in the Eqn. 3 are separately 1,1,1,4 from $\mathbf{X}^{(1)}$ to $\mathbf{X}^{(4)}$. We adopt SGD optimizer with momentum = 0.9, and employ the "poly" learning rate schedule policy with initial learning rate = 0.001, weight decay = 0.001. New convolution layers are initialized by the kaiming normalization [14]. On the PASCAL VOC 2012 dataset, we set the balance factor λ_1 in the Eqn. 11 to 100 while on the PASCAL Context dataset, this value is set to 1 since there are more semantic labels where a large balance factor will lead to over-smooth, impacting the segmentation performance. The training images are augmented by random scaling ($[0.5, 2.0]$), random flipping ($p = 0.5$), random rotating ($[-10, 10]$), and are randomly cropped into size 512×512. We run 200 training epochs on a single NVIDIA TitanX 1080ti GPU. When producing the segmentation results on the validation set, we utilize the multi-scale and flipping techniques. Note that we do not employ CRF [21] post-processing in all experiments.

Method	Supervision	Backbone	mIoU
SEC [20]	I	VGG16	50.7
AugFeed [32]	I	VGG16	54.3
STC [40]	I	VGG16	49.8
AffinityNet [2]	I	ResNet38	58.4
GAIN [24]	I	VGG16	55.3
MDC [41]	I	VGG16	60.4
SeeNet [16]	I	VGG16	61.1
FickleNet [23]	I	ResNet101	61.2
SSNet [44]	I	VGG16	57.1
OAA [18]	I	VGG16	63.1
ICD [9]	I	VGG16	64.0
BoxSup [7]	B	VGG16	62.0
WSSL [30]	B	VGG16	60.6
SDI [19]	B	VGG16	65.7
ScribbleSup* [26]	S	VGG16	63.1
RAWKS [37]	S	ResNet101	59.5
NormalCut [34]	S	ResNet101	72.8
KernelCut [35]	S	ResNet101	73.0
BPG [38]	S	ResNet101	73.2
PSI(ours)	S	ResNet101	**74.9**

Table 1. Comparison with state-of-the-art methods on the PASCAL VOC 2012 validation set. 'I' means the image level tags, 'B' means boxes and 'S' means scribbles. The symbol '*' means the segmentation predictions are post-processed by the CRF.

5.3. Comparison with state-of-the-art methods

PASCAL VOC 2012. We compare the segmentation performance of our proposed PSI with other state-of-the-art methods on the PASCAL VOC 2012 dataset [8]. The de-

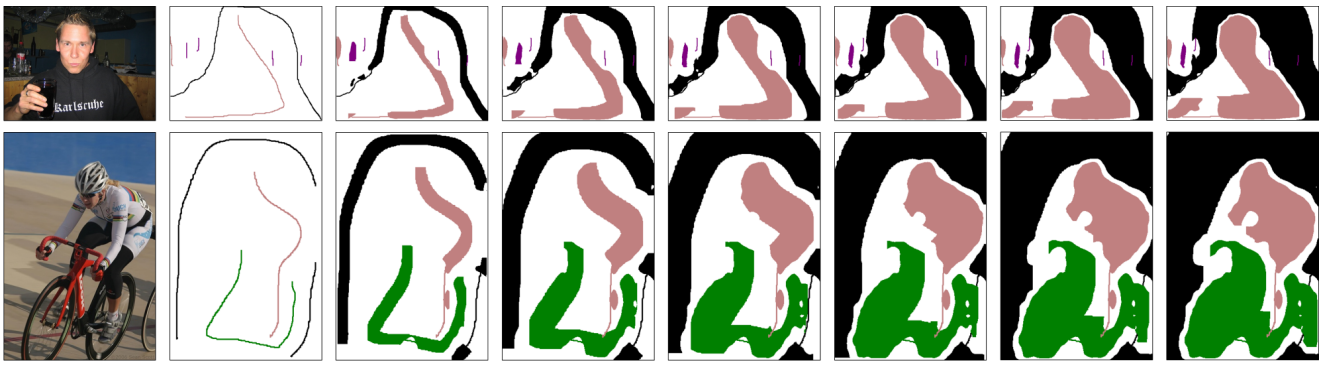

Figure 2. Evolved supervisions progressively expanded by SLD with different iteration times.

tailed results are reported in the Table 1. As can be seen, our proposed PSI achieves the best segmentation performance over all compared state-of-the-art weakly supervised segmentation methods. Our method, as well as most scribble supervised methods, performs better than the image-tag and box supervised methods due to the superiority of scribble supervision. As to other scribble supervised methods which employ the same backbone structure, our method is able to reach an improvement of 2.1%, 1.9% and 1.7% over NormalCut [34], KernelCut [35] and BPG [38] without CRF [21] post-processing. Note that NormalCut [34] and KernelCut [35] specifically designed losses which require specific computation of the bilateral filtering while our method only uses the simple cross entropy and $\mathcal{L}_1$ losses. BPG [38] introduced auxiliary edge information while our method only uses the initial scribbles. Despite the concision of our framework, we also achieve the best result, demonstrating the effectiveness of our proposed PSI framework.

Method	Supervision	mIoU(%)
ScribbleSup* [26]	Scribble	36.1
RAWKS [37]	Scribble	36.0
BoxSup [7]	Semi	40.5
PSPNet [46]	Full	47.8
DANet [10]	Full	52.6
OCR [43]	Full	56.2
DeepLabv3+	Scribble	37.1
PSI(ours)	Scribble	**43.1**

Table 2. Comparison with state-of-the-art methods on the PASCAL Context dataset following a pure weakly supervised setting.

PASCAL Context. We further conduct experiments on the PASCAL Context dataset [29] to verify the generalization capability of our proposed method. The comparison results are reported in the Table 2. ScribbleSup [26] and RAWKS [37] have publicly reported their results following the pure weakly supervised setting. The results reported by BoxSup [7] employed the fully annotated masks in the PASCAL Context dataset as well as weakly labeled boxes in the PASCAL VOC 2007 dataset [8]. To fairly compare our result, we also design a baseline method which simply utilizes DeepLabV3+ [6]. As can be seen from Table 2, our proposed PSI boosts the baseline result by 6.0% to 43.1% which outperforms all weakly and semi-supervised methods. Considering the difficulty of the PASCAL Context dataset [29], the obviously non-trivial improvement illustrates the robustness of our proposed PSI framework. It could also be observed that PSI could greatly narrow the performance gaps between the weakly supervised methods and the fully supervised methods. All the above analyses further verify the effectiveness of our proposed progressive segmentation inference framework.

5.4. Ablation studies

Effectiveness of different components. We conduct ablation studies to verify the effectiveness of our proposed CPP and SLD module. The detailed results are reported in the Table 3 and the Table 4. All the experiments are conducted under the single-scale testing setting. we take retrained DeepLabV3+ [6] as our baseline and it can achieve a segmentation accuracy of 66.6%. Only adding the CPP module could lead to a performance of 70.1%, obtaining a significant improvement of 3.5% which clearly demonstrates the effectiveness of CPP. Besides, we adopt other pattern aggregation modules including UP which fuses patterns with the bilinear unsampling and convlutional layers as well as SF [25] which aligns patterns with Semantic Flow. Our CPP module improves by 2.3%, 1.6% in comparison to UP and SF (74.9% vs 72.6%, 73.3%). It further indicates the effectiveness of the multi-granular pattern aggregation in our CPP. If only the SLD module is implemented, the result reaches 69.4% and the improvement is 2.8%. It is obviously a non-trivial boost, indicating the power of our proposed progressive semantic label inference process. Appending the SLD module to the CPP module

obtains a further performance promotion of 2.6%, showing that our proposed CPP and SLD module are compatible and could cooperate with each other to lead to a satisfactory performance. Further applying a multi-scale testing boosts the performance to 74.9%, setting a new state-ot-the-art.

CPP		✓		✓	✓
SLD			✓	✓	✓
Multi-scale					✓
mIoU(%)	66.6	70.1	69.4	72.9	**74.9**

Table 3. Comparisons of our method using different components on the PASCAL VOC 2012 validation set.

aggregation module	UP	SF	CPP
mIoU(%)	72.6	73.3	**74.9**

Table 4. Performance comparisons with different pattern aggregation modules on the PASCAL VOC 2012 validation set.

Destination Layer. In CPP, one key issue is to choose which layer to serve the destination layer. We choose four layers (Conv2-Conv5 from the encoder) as the candidate destination layer and denote them separately as $\mathbf{X}^{(4)}, \mathbf{X}^{(3)}, \mathbf{X}^{(2)}, \mathbf{X}^{(1)}$. The detailed results are reported in the Table 5. It can be seen that with the semantic level of the destination layer increasing, the segmentation performance keeps going up. It means that high-level semantic information devotes more for the learning of segmentation model, and with the supplements of low-level visions information, the feature representations are more robust. It also indicates that the high-level semantics of the destination layer matters more for the CPP module. The $\mathbf{X}^{(1)}$ variant integrates the multi-scale contextual patterns to the $8\times$ resolution layer while the baseline DeepLabV3+ uses a $4\times$ layer to aggregate the information. The consistent performance improvement from $\mathbf{X}^{(1)}$ to $\mathbf{X}^{(4)}$ demonstrates that our proposed CPP module is able to deal with the potential appearance noise and consistently aggregate the information at various semantic levels, making up the shortages of contextual patterns scarcity.

k	1	2	3	4
mIoU(%)	65.33	65.58	68.84	**70.09**

Table 5. Performance comparisons of our method using different destination layers on the PASCAL VOC 2012 validation set.

The number of SLD steps. In SLD, one key variant to determine is the update step number t. To decide a proper t, we here conduct a series of experiments where the step number t increases from 1 to 6, as shown in the Table 6. The segmentation accuracy raises with the SLD step increasing in the beginning when t ranges from 1 to 3. It means that our method is able to absorb more useful label information for the network training to circumvent the scarcity of original supervisions, verifying the effectiveness of the proposed SLD module. Afterwards, the segmentation performance saturates at $t = 3, 4$ and slightly drops when $t = 5, 6$. Since a too large step t expands supervisions to a broad unlabeled region, the imperfection of the depended segmentation predictions will introduce unexpectable noises, affecting the performance. We have also displayed the evolved supervisions progressively generated by SLD in the Fig. 2. As it shows, with the step t increasing, the expand regions grow and become more dependent on the confidence map of the segmentation predictions. Based on the above results, we adopt the middle step $t = 3$ to be a proper SLD step.

iter_time	1	2	3	4	5	6
mIoU(%)	72.50	72.85	**72.88**	**72.88**	72.75	72.74

Table 6. Comparisons of our method using different iteration time on the PASCAL VOC 2012 validation set.

6. Conclusion

To address scribble-supervised semantic segmentation, a progressive segmentation inference (PSI) framework is proposed. In PSI, we specifically develop two crucial modules, contextual pattern propagation (CPP) and semantic label diffusion (SLD) to enhance and refine pixel-level segmentation results from partially known seeds. CPP effectively integrates the patterns of multiple granularities as well as different locations through graphical model, to aid the inference of the segmentation results. In addition, SLD is developed to progressively expand the supervisions through a dynamically learned decision strategy. CPP and SLD are finally modularized into a close-looping between the segmentation network update and supervision refinement. Extensive experiments have proved the effectiveness of our proposed CPP and SLD modules. Meantime, our proposed PSI achieves the state-of-the-art performances on the challenging PASCAL VOC 2012 and PASCAL Context datasets.

7. Acknowledgement

This work was supported by the Natural Science Foundation of Jiangsu Province (Grant No. BK20190019), the National Natural Science Foundation of China (Grants Nos. 62072244, 61972204), the Fundamental Research Funds for the Central Universities (No. 30921011104), the Natural Science Foundation of Shandong Province (Grant No. ZR2020LZH008), and partly collaborated with State Key Laboratory of High-end Server & Storage Technology.

References

[1] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, volume 29, pages 753–762. Wiley Online Library, 2010. 2

[2] 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*, pages 4981–4990, 2018. 6

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[5] 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*, 40(4):834–848, 2017. 1, 2

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 4, 6, 7

[7] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 1, 6, 7

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6, 7

[9] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 1, 6

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 1, 2, 7

[11] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008. 2

[12] Leo Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1768–1783, 2006. 1, 2

[13] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018. 6

[17] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 2

[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*, pages 2070–2079, 2019. 1, 6

[19] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1, 6

[20] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 6

[21] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 2, 6, 7

[22] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 2

[23] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5267–5276, 2019. 6

[24] 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*, pages 9215–9223, 2018. 1, 6

[25] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *European Conference on Computer Vision*, pages 775–793. Springer, 2020. 7

[26] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for

semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 1, 2, 6, 7

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 6, 7

[30] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 1, 6

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 6

[32] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *European conference on computer vision*, pages 90–105. Springer, 2016. 6

[33] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 1, 2

[34] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018. 1, 3, 6, 7

[35] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 1, 3, 6, 7

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[37] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1, 2, 6, 7

[38] Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In *IJCAI*, pages 3663–3669, 2019. 1, 2, 6, 7

[39] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[40] 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*, 39(11):2314–2320, 2016. 6

[41] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 1, 6

[42] Chunyan Xu, Li Wei, Zhen Cui, Tong Zhang, and Jian Yang. Meta-vos: Learning to adapt online target-specific segmentation. *IEEE Transactions on Image Processing*, 30:4760–4772, 2021. 2

[43] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 1, 2, 7

[44] 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*, pages 7223–7233, 2019. 6

[45] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 2

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 7

[47] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4514–4523, 2020. 3