Saliency Guided Self-Attention Network for Weakly and Semi-Supervised Semantic Segmentation

QI YAO and XIAOJIN GONG, (Member, IEEE)
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
Corresponding author: Xiaojin Gong (gongxj@zju.edu.cn)

ABSTRACT Weakly supervised semantic segmentation (WSSS) using only image-level labels can greatly reduce the annotation cost and therefore has attracted considerable research interest. However, its performance is still inferior to the fully supervised counterparts. To mitigate the performance gap, we propose a saliency guided self-attention network (SGAN) to address the WSSS problem. The introduced self-attention mechanism is able to capture rich and extensive contextual information but may mis-spread attentions to unexpected regions. In order to enable this mechanism to work effectively under weak supervision, we integrate class-agnostic saliency priors into the self-attention mechanism and utilize class-specific attention cues as an additional supervision for SGAN. Our SGAN is able to produce dense and accurate localization cues so that the segmentation performance is boosted. Moreover, by simply replacing the additional supervisions with partially labeled ground-truth, SGAN works effectively for semi-supervised semantic segmentation as well. Experiments on the PASCAL VOC 2012 and COCO datasets show that our approach outperforms all other state-of-the-art methods in both weakly and semi-supervised settings.

INDEX TERMS Weakly and semi-supervised semantic segmentation, self-attention, saliency.

I. INTRODUCTION

Semantic segmentation aims to predict a semantic label for each pixel in an image. Based upon the fundamental Fully Convolutional Networks (FCNs) [1], various techniques such as dilated convolution [2], spatial pyramid pooling [3], and encoder-decoders [4] have been developed in the last decade. These techniques gradually improve segmentation accuracy via exploiting extensive contextual information. Recently, the self-attention mechanism [5]–[7] has been successfully employed to capture richer contextual information and boost the segmentation performance further. Although the above-mentioned methods have achieved high performance in semantic segmentation, they all work under full supervision. This supervision manner requires a large amount of pixel-wise annotations for training, which are very expensive and time-consuming.

To reduce the annotation burden, different supervision forms such as bounding boxes [8], scribbles [9], and image-level tags [10] have been considered for semantic segmentation. Among them, the form of using image-level tags has attracted major attention because of its minimal annotation cost as well as its great challenge. Recent work [11] has shown that convolutional neural networks (CNNs) have the localization ability even if only image-level tags are used. This observation has inspired many weakly-supervised semantic segmentation (WSSS) researches. However, attentions in the class activation maps (CAMs) [11] inferred from image classification networks tend to focus on small discriminative parts of objects. The object location cues (also referred to as seeds) retrieved from these CAMs are too sparse to effectively train a segmentation model. Therefore, a great amount of effort has been devoted to recover dense and reliable seeds [12]–[16].

In this paper, we aim to take advantage of the self-attention mechanism to mine high-quality seeds. As validated in [5], [6], this mechanism is able to successfully capture long-range contextual dependencies in fully-supervised semantic segmentation. However, it encounters the following challenges when applied to WSSS. (1) Some foreground objects may always be co-occurrent with the same background, like ‘boat’ and ‘water’, leading to a pathological
bias [17]; (2) The global average pooling (GAP), which is commonly used in classification networks to aggregate pixel-wise responses into image-level label scores, encourages all responses to be high; (3) In the self-attention scheme, each pixel directly contributes to all other pixels and vice versa. These factors may result in a mis-spread of attentions from discriminative parts to unexpected regions. Typical examples are illustrated in Figure 1. In the fully-supervised setting, the information of the selected discriminative pixels can be correctly propagated to the pixels belonging to the same category. Contrastively, the discriminative information is diffused to the regions of the background and other categories under weak supervision.

To address the above-mentioned problems and enable the self-attention mechanism to capture long-range contextual information correctly under weak supervision, we construct a self-attention network that leverages the class-agnostic saliency as a guidance. A saliency map provides a rough detection of foreground objects so that it can prevent attentions from spreading to unexpected background regions. To further reduce information diffusion among foreground categories, we integrate the class-specific attention cues as additional supervision. The integration of these prior cues is implemented in our network via a joint learning of a seed segmentation branch and an image classification branch. By all means, our network generates high quality seeds so that the segmentation performance is boosted.

- In our network, both the seed segmentation branch and the image classification branch can produce high quality seed results. The ensemble of two results improves the quality of seeds further.
- By simply replacing saliency maps and attention cues with partially labeled segmentation ground-truth, SGAN can work effectively for semi-supervised semantic segmentation as well.
- Our approach achieves state-of-the-art performance on the PASCAL VOC 2012 and COCO datasets in both weakly and semi-supervised settings.

II. RELATED WORK

A. WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Various supervision forms have been exploited for weakly-supervised semantic segmentation (WSSS). Here, we focus on the works using image-level tags. Most recent methods solve the WSSS problem by first mining reliable seeds and then take them as proxy ground-truth to train segmentation models. Thus, a great amount of effort has been devoted to generate high-quality seeds.

A group of approaches take the class activation maps (CAMs) [11] generated from classification networks as initial seeds. Since CAMs only focus on small discriminative regions which are too sparse to effectively supervise a segmentation model, various techniques such as adversarial erasing [12], [17]–[19], attention accumulation [20], [21], and region growing [13], [22], [23] have been developed to expand sparse object seeds. Another research line introduces dilated convolutions of different rates [14]–[16], [24] to enlarge receptive fields in classification networks and aggregate multiple attention maps to achieve dense localization cues. In this work, we adopt the self-attention mechanism to capture richer and more extensive contextual information in order to mine high quality seeds.

B. SELF-ATTENTION MECHANISM

The self-attention mechanism [25] computes the context at each position as a weighted sum of all positions. Its superiority in capturing long-range dependencies has been recently validated by various vision tasks [5], [6], [26], [27]. Particularly, in semantic segmentation, Yuan and Wang [5] integrated this mechanism into pyramid structures to capture multi-scale contextual information; Fu et al. [6] constructed a dual attention network to capture dependencies in both spatial and channel dimensions; Huang et al. [28] proposed an interlaced sparse approach to improve the efficiency of the self-attention mechanism; and Huang et al. [7] designed a recurrent criss-cross attention module to efficiently harvest the contextual information. These methods significantly boost the segmentation performance, but all of them perform under full supervision. Although Sun and Li [22] utilized the self-attention scheme for WSSS, they only used this scheme to learn a saliency detector that is trained also in a fully-supervised manner. In our work, we apply the
self-attention scheme to a weakly-supervised scenario which is more challenging.

C. SALIENCY GUIDANCE FOR WSSS
Salient object detection (SOD) [29] produces class-agnostic saliency maps that distinguish foreground objects from the background. The results of SOD have been extensively used in weakly-supervised semantic segmentation. For instance, many methods [12]–[15], [17], [24] exploited saliency maps to generate background seeds. Moreover, Wei et al. [30] adopted a self-paced learning strategy to learn a segmentation model that was initialized under the full supervision of saliency maps of simple images. Sun and Li [22] utilized saliency maps to guide a CAM-seeded region growing process to expand object regions. Fan et al. [31] used instance-level saliency maps to construct and partition similarity graphs for WSSS. In addition, Chaudhry et al. [19], Oh et al. [32], and Wang et al. [33] combined class-agnostic saliency maps with class-specific attention cues like us to obtain dense seed. But their combinations are implemented in user-defined ways. In contrast, our saliency maps and attention cues are adaptively fused within the proposed self-attention network.

III. THE PROPOSED APPROACH
The proposed approach for weakly and semi-supervised semantic segmentation is divided into two parts: (1) learning a saliency guided self-attention network to generate dense and accurate seeds, and (2) utilizing the high-quality seeds as proxy ground-truth to train a semantic segmentation model. The details are introduced in the followings.

A. SALIENCY GUIDED SELF-ATTENTION NETWORK
1) NETWORK ARCHITECTURE
The overview of our proposed saliency guided self-attention network (SGAN) is illustrated in Figure 2. It consists of three components: (1) a CNN backbone to learn deep feature representations; (2) a saliency guided self-attention module that propagates attentions from small discriminative parts to non-discriminative regions via capturing long-range contextual dependencies; (3) an image classification branch together with a seed segmentation branch to supervise the training of the entire network.

We adopt a slightly modified VGG-16 [10] network as the backbone for feature extraction. The last two pooling layers are removed in order to increase the resolution of the output feature maps. Note that, unlike previous works [14], [15], [24] that enlarge the dilation rate of convolution kernels in conv5 block, we avoid the usage of dilated convolution and instead use the self-attention module to capture more extensive contexts.

2) SALIENCY GUIDED SELF-ATTENTION MODULE
This module aims to take advantage of the self-attention mechanism to capture rich contextual information that is essential for discovering integral extent of objects and retrieving high-quality seeds. The self-attention mechanism has demonstrated its effectiveness in capturing long-range dependencies under full supervision [5], [6]. However, simply integrating it into a weakly-supervised network may suffer from a severe mis-spread problem as introduced in Section I. Thus, we propose to incorporate class-agnostic saliency priors to prohibit the spread of attentions from discriminative object regions to the background.

We formally describe the saliency guided self-attention module as follows. This module takes the feature map output from the VGG’s conv5 block, which is denoted as $X \in \mathbb{R}^{C \times H \times W}$, together with a saliency map as the inputs. With the input feature map, a sequence of spatial matrix operations are performed to generate a spatial attention map $P \in \mathbb{R}^{N \times N}$, where $N = H \times W$ is the number of positions. More specifically, $X$ is first fed into two $1 \times 1$ convolutions for...
linear embedding and generating a key feature map \( K \in \mathbb{R}^{C \times H \times W} \) and a query feature map \( Q \in \mathbb{R}^{C \times H \times W} \) respectively. These two feature maps are further reshaped to \( \mathbb{R}^{C \times N} \). Then, the spatial attention map \( P \) is generated by computing the inner product of channel-wise features from any two positions of \( K \) and \( Q \). That is,

\[
P_{ij} = K^T_{ij}Q_{ij},
\]

(1)

where \( i, j \in \{1, 2, \ldots, N\} \) are the indexes of positions, and \( K_{ij}, Q_{ij} \in \mathbb{R}^{C \times 1} \) are the channel-wise features. \( P_{ij} \) measures the similarity of the features extracted at position \( i \) and \( j \). Note that different pairwise functions [26] can be used for the similarity measurement, we take the inner product because it is simple but effective enough.

For the input saliency map, we first threshold it to get a binary mask \( B \) and reshape it to \( \mathbb{R}^{N \times 1} \). After that, a saliency attention map \( S \in \mathbb{R}^{N \times N} \) is computed by

\[
S_{ij} = \mathbb{1}(B_i \neq B_j),
\]

(2)

where \( \mathbb{1} \) is an indicator function. It equals one if both positions \( i \) and \( j \) are salient or non-salient.

Then, the context attention map \( D \in \mathbb{R}^{N \times N} \) is generated via an element-wise production between the spatial attention map \( P \) and the saliency attention map \( S \), followed by a linear normalization:

\[
D_{ij} = \frac{P_{ij} \cdot S_{ij}}{\sum_{j=1}^{N} P_{ij} \cdot S_{ij}}.
\]

(3)

Once the context attention map \( D \) is obtained, we use it to enhance the original feature map \( X \). Specifically, we reshape \( X \) to \( \mathbb{R}^{C \times N} \) and conduct a matrix multiplication between \( X \) and the transpose of \( D \). Then we reshape the result back to \( \mathbb{R}^{C \times H \times W} \) and perform an element-wise summation with \( X \) to obtain the enhanced features \( E \in \mathbb{R}^{C \times H \times W} \). That is,

\[
E_i = \gamma \sum_{j=1}^{N} (D_{ij} \cdot X_j) + X_i,
\]

(4)

where \( \gamma \) is a parameter initialized as 0 [6] and gradually learned in training. Equation (4) indicates that each position of \( E \) is the sum of similarity-weighted features at all positions and the original features. Therefore, this module captures contextual information from a whole image. By this means, attentions on small discriminative parts of objects can be propagated to non-discriminative object regions, but not to the background because of the guidance of saliency.

3) INTEGRATING CLASS-SPECIFIC ATTENTION CUES

The class-agnostic saliency maps introduced above can only roughly separate foreground objects from the background, but provide no information about semantic categories. In order to prevent our SGAN from mis-spreading attentions among objects of different categories, we propose to integrate the class-specific attention cues obtained by the CAM method [11] from a classification network as additional supervision.

Specifically, we construct a segmentation branch in our SGAN. It takes the enhanced feature \( E \) as the input and goes through a convolutional layer to produce \( M \) segmentation maps, each of which corresponds to a foreground category. Meanwhile, we retrieve reliable but sparse foreground object seeds by thresholding the class activation maps obtained from the VGG-16 classification network with a high confidence value (empirically set to 0.3 in this work) and use them to supervise the segmentation maps. The corresponding seed loss \( L_{seed} \) is defined by

\[
L_{seed} = \frac{1}{\sum_{z \in Z} |A_z|} \sum_{z \in Z} \sum_{u \in A_z} \log \Phi_{z,u}.
\]

(5)

Here, \( Z \) denotes the set of foreground classes that present in an image and \( A_z \) is a set of seed locations corresponding to class \( z \). \( | \cdot | \) is the cardinality of the set. \( \Phi_{z,u} \) denotes the probability of class \( z \) at position \( u \) in the segmentation maps. Note that, in contrast to the seeding loss defined in [10], [13] that considers both foreground and background categories, our loss only takes into account the foreground classes.

4) TRAINING SGAN

The network also has an image classification branch that is supervised by image-level labels. Let us denote the classification probability as \( \tau \in \mathbb{R}^{M \times 1} \) and the corresponding image-level label as \( y = [y_1, \ldots, y_M] \in \{1, -1\} \), which indicates the presence or absence of foreground object categories. Then the classification loss \( L_{cls} \) is defined by the sigmoid cross entropy. That is,

\[
L_{cls} = -\frac{1}{M} \sum_{m=1}^{M} \log (y_m \cdot (\tau_m - \frac{1}{2}) + \frac{1}{2}).
\]

(6)

The overall loss for training our saliency guided self-attention network is defined by

\[
L = L_{cls} + \lambda L_{seed},
\]

(7)

where \( \lambda \) is a weighting factor to balance the two terms.

5) GENERATING HIGH-QUALITY SEEDS

Once the proposed SGAN is trained, we note that there are two possible ways to get class activation maps (CAMs). One is following the common practice [11] to infer the CAMs from the image classification branch. The other is directly taking the segmentation maps output from the seed segmentation branch. Either may be used to retrieve dense and accurate seeds. But we find out that the combination of them improves the seeds’ quality further because they are complementary in some scenarios as will be shown in Section IV-D. Therefore, we take an element-wise summation of these two results to generate the final class activation maps.

Then, for each foreground class, we retrieve object seeds by thresholding the corresponding class activation map with a high value \( \alpha \). In addition, we retrieve background seeds by thresholding the input saliency map with a low value \( \beta \). Following [10], [13], [14], we set \( \alpha = 0.2 \) and \( \beta = 0.06 \) in our experiments.
B. TRAINING THE SEGMENTATION NETWORK

After obtaining the high-quality seeds, we can use them as proxy ground-truth labels to train an arbitrary semantic segmentation network. In this work, we adopt the balanced seed loss $L_{\text{balance seed}}$ proposed in DSRG [13] for the seed supervision. It is

$$L_{\text{balance seed}} = \frac{1}{|\mathcal{A}_z|} \sum_{z \in \mathcal{Z}} \sum_{u \in \mathcal{A}_z} \log \Phi_z(u) - \frac{1}{|\mathcal{A}_z|} \sum_{z \in \mathcal{Z}} \sum_{u \in \mathcal{A}_z} \log \Phi_z(u),$$

(8)

where $\mathcal{Z}$ denotes the background. $\mathcal{A}_z$, $\mathcal{A}_z$, and $\Phi_z(u)$ holds the same definitions as previous.

We further exploit the boundary constraint loss used in both DSRG [13] and SEC [10] to encourage segmentation results to match up with object boundaries. Let us denote $\mathbf{I}$ as the input image and $\mathbf{R}(\mathbf{I}, \Phi)$ as the output probability map of the fully-connected CRF. Then the boundary constraint loss is defined as the mean KL-divergence between the segmentation map and the output of CRF:

$$L_{\text{boundary}} = \frac{1}{N} \sum_{u=1}^{M+1} \sum_{z=1}^{N} \mathbf{R}_{z,u}(\mathbf{I}, \Phi) \log \frac{\mathbf{R}_{z,u}(\mathbf{I}, \Phi)}{\Phi_{z,u}}.$$ (9)

Thus, the total loss for training the segmentation model is $L = L_{\text{balance seed}} + L_{\text{boundary}}$.

C. SGAN UNDER SEMI-SUPERVISION

In the semi-supervised setting, a small number of training images are provided with strong pixel-level labels and the rest have image-level tags only. For these strongly annotated images, we replace their saliency maps with the binary foreground masks derived from the ground-truth annotations, and meanwhile use the ground-truth to supervise the seed segmentation branch. These simple replacements help the proposed SGAN to learn a better model and generate higher-quality seeds for all images.

For the training of the segmentation network, we also have the ground-truth to take place of the generated seeds for these strongly annotated images. Except this, we keep the training loss under semi-supervision the same as that in the weakly-supervised setting.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASET AND EVALUATION METRIC

The proposed approach is evaluated on the PASCAL VOC 2012 segmentation benchmark [34] and the COCO [35] dataset.

$a$: PASCAL VOC

This dataset provides pixel-wise annotations for 20 object classes and one background class. It contains 1464 images for training, 1449 images for validation and 1456 images for testing. Following the common practice [10], [13], [14], we augment the training set to 10,582 images. Our network is trained on the augmented training set only using image-level annotations and evaluated on the validation and test set in terms of the mean intersection-over-union (mIoU) criterion.

$b$: COCO

This dataset contains 80k images for training and 40k images for validation. Our network is trained on the training set with only image-level annotations and evaluated on the validation set in terms of mIoU for 81 categories.

2) TRAINING DETAILS

The saliency guided self-attention network is built on the VGG-16 network pre-trained on ImageNet. The remaining parameters of our SGAN are randomly initialized. Following [14], we use S-Net [29] to get a class-agnostic saliency map for each image. SGD with mini-batch is used for training. The batch size is set to 15, the momentum is 0.9 and the weight decay is 0.0005. Input images are resized to 321 × 321 and no data augmentation except randomly horizontal flip are adopted. We train the SGAN for 8,000 iterations. The initial learning rate is 0.001 and it is decreased by a factor of 10 every 2,000 iterations.

The semantic segmentation model is chosen to use the Deeplab-ASPP [2] network in order to compare with other WSSS works. Both VGG-16 and ResNet-101 backbones are investigated. The batch size is set to 15, the momentum is 0.9 and the weight decay is 0.0005. Input images are resized to 353 × 353 and randomly cropped to 321 × 321 for training. Horizontal flip and color jittering are employed for data augmentation. We train the segmentation model for 12,000 iterations. The initial learning rate is 0.001 and it is decreased by a factor of 0.33 every 2,000 iterations.

3) REPRODUCIBILITY

We implement our SGAN on PyTorch [36] for training and producing high-quality seeds. We use the official Deeplab-ASPP code implemented on Caffe [37] for semantic segmentation. All the experiments are conducted on a GTX 1080Ti GPU. The code is available at https://github.com/yaoqi-zd/SGAN.

B. COMPARISON TO THE STATE OF THE ART

1) WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

$a$: PASCAL VOC

We compare our approach with other state-of-the-art WSSS methods that are also supervised only by image-level labels. For fair comparison, we separate the methods into two groups according to the backbones upon which their segmentation models are built, as listed in Table 1. Most of existing methods use saliency maps to retrieve background seeds or even foreground seeds. Therefore, Table 1 also marks out if a method uses saliency maps.

Table 1 shows that our method outperforms all the previous methods on both VGG-16 and ResNet-101 backbones.
In particular, AE_PSL [12], GAIN [17], and SeeNet [18] use erasing techniques to get dense localization cues, which tend to identify true negative regions. AffinityNet [40], DSRG [13], SGDN [22], and SSD[23] adopt region-growing techniques to expand seeds. It may be hard for them to expand to non-discriminative regions if initial seeds are concentrated on extremely small discriminative parts. OAA [20] accumulates attention maps during the training procedure which may introduce unexpected attention regions at the early stage when the classifier is weak. MDC [14], DFPN [16], and FickleNet [15] use dilated convolutions to retrieve dense seeds, whose receptive fields are not adaptive to image contents and may result in over-expansion. In contrast, our method can achieve dense and accurate seeds, which is beneficial from the self-attention mechanism as well as the design of our SGAN network.

It needs to be mentioned that our approach outperforms DSNA [41], which uses a spatial attention mechanism, and SSNet [42] that jointly learn saliency and segmentation with additional pixel-wise saliency supervision.

b: COCO

To further validate the effectiveness of our approach, we conduct experiments on the COCO dataset which is much more challenging than PASCAL VOC. Most existing methods haven’t done the experiments on COCO yet, except DSRG [13]. Therefore, we compare our results to DSRG in Table 2, in which the results of SEC are also quoted from DSRG’s paper. Table 2 shows that our VGG16-based SGAN surpasses both SEC and DSRG by a large margin. In particular, our method performs excellently on those categories with large scale, such as airplane, bus and train etc, but has difficulties in handling small things that are likely to be indistinguishable with the clustered background, such as baseball glove and spoon etc.

2) SEMI-SUPERVISED SEMANTIC SEGMENTATION

We compare our approach with other state-of-the-art semi-supervised semantic segmentation methods on PASCAL VOC 2012 dataset under the same setting, where 1.4K
TABLE 3. Comparison of semi-supervised semantic segmentation methods on PASCAL VOC 2012 validation set in terms of mIoU. All methods are based on the VGG-16 backbone.

| Methods          | Publication | Training Set  | mIoU   |
|------------------|-------------|---------------|--------|
| Deeplab [2]      | TPAMI'17    | 1.4K strong   | 62.5   |
|                  |             | 10.6K strong  | 70.3   |
| WSSL [44]        | ICCV'15     | 1.4K strong + 9K weak | 64.6   |
| GAIN [17]        | CVPR'18     | 1.4K strong + 9K weak | 60.5   |
| MDC [14]         | CVPR'18     | 1.4K strong + 9K weak | 65.7   |
| DSRG [13]        | CVPR'18     | 1.4K strong + 9K weak | 64.3   |
| PickleNet [15]   | CVPR'19     | 1.4K strong + 9K weak | 65.8   |
| SGAN(Ours)       | -           | 1.4K strong + 9K weak | 67.0   |

images annotated with pixel-level labels and 9K images annotated with image-level tags are available. The comparison results are reported in Table 3, together with the results obtained by the fully supervised Deeplab-ASPP [2] segmentation network. Table 3 shows that our approach not only outperforms all previous methods but also reaches 95.3% of the performance under full supervision. It needs to be mentioned that most previous methods use the 1.4K pixel-level labels for training the semantic segmentation model only, while our approach can easily adopt them to facilitate the training of SGAN and improve the quality of dense seeds for the 9K weakly annotated images, leading to better performance.

C. QUALITATIVE RESULTS

Figure 3 shows qualitative segmentation results obtained by the proposed approach on PASCAL VOC dataset. As we can see, our approach produces accurate segmentation results and recovers fine details of object boundaries for images containing scale variation, multiple objects, and complex background. A typical failure case is also presented in the last row, in which the dining table is indistinguishable from the background and thus misidentified as background.

D. ABLATION STUDIES

1) EFFECTIVENESS OF THE COMPONENTS IN SGAN

To investigate the effectiveness of each component in the saliency guided self-attention network, we conduct a series of experiments in different settings while keeping the VGG-16-based segmentation model the same throughout all these experiments. Particularly, the following configurations are investigated: (1) the full model, which is referred to as SGAN; (2) the full model but only using the CAMs from the image classification branch, which is denoted as SGAN\textsuperscript{CLS}; (3) the full model but only using the CAMs from the seed segmentation branch, which is denoted as SGAN\textsuperscript{SEG}; (4) the model without the segmentation branch and the seed loss, which is denoted as SGAN-SEED; (5) the model without the segmentation branch, the seed loss, and the saliency guidance, which in essence is directly integrating the self-attention mechanism into the modified VGG-16 classification network. We denote this variant as SGAN-SAL-SEED; and (6) the baseline model without our proposed saliency guided self-attention module, which is actually the modified VGG-16 classification network.

The comparison results are listed in Table 4, from which we make the following observations: (1) The SGAN-SAL-SEED model that applies the self-attention mechanism directly to a weakly-supervised network degrades the segmentation performance, especially for the categories that are always co-occurrent with the same background, for instance, ‘airplane’ with ‘sky’, ‘boat’ with ‘water’, ‘horse’ with ‘grass’, etc. In such cases, SGAN-SAL-SEED tends to propagate attentions from foreground objects to the concurrent background and generate inaccurate seeds. (2) The SGAN-SEED model that uses the proposed saliency guided self-attention module outperforms the baseline model over all categories. (3) After integrating class-specific attention cues, both SGAN\textsuperscript{CLS} and SGAN\textsuperscript{SEG} produce high quality seeds and achieve high segmentation performance. The ensemble of CAMs from two branches, that is SGAN, boosts the performance further. (4) Compared to the baseline model, our full model boosts the performance significantly for the categories containing objects in large size, like ‘bus’ (+29.2%) and ‘train’ (+16.5%), and the categories with large scale variation such as ‘person’ (+14%). For these categories, the initial localization cues are usually too sparse.
TABLE 4. Comparison of the proposed model under different settings on VOC 2012 val set in terms of mIoU (%).

| Method     | mIoU |
|------------|------|
| Baseline   | 55.0 |
| SGAN-SAL-SEED | 45.8 |
| SGAN-SEED  | 62.4 |
| SGAN<sup>CLS</sup> | 63.7 |
| SGAN<sup>SEG</sup> | 63.4 |
| SGAN       | 64.2 |

![Comparison of the proposed model under different settings on VOC 2012 val set in terms of mIoU (%).](image)

FIGURE 4. Visualization of the class attention maps learned by different variants of our SGAN.

to delineate the integral object extent. Our model can effectively propagate attentions from small discriminative parts to non-discriminative regions of objects and generate more complete object seeds, leading to much better segmentation performance.

In order to understand these models more intuitively, we present the class activation maps generated by each variant in Figure 4. From this figure we observe that SGAN-SAL-SEED tends to diffuse class-specific attentions to backgrounds and other categories. SGAN-SEED can greatly reduce the diffusion to backgrounds but it cannot prevent the miss-spread among foreground categories. SGAN<sup>CLS</sup>, SGAN<sup>SEG</sup>, and SGAN can constrain attentions mostly within the regions of the right class. Moreover, the CAMs obtained by SGAN<sup>CLS</sup> and SGAN<sup>SEG</sup> are complementary to each other in some cases such as the ‘train’ image and the ‘sheep’ image.

2) EVALUATION OF SEEDS’ QUALITY

High-quality seeds are obtained by thresholding the class activation maps presented above. Here, we also adopt precision, recall, and the F-measure score to evaluate the quality
The quality of seeds generated by different variants of our SGAN model.

| Method      | Precision | Recall | F1-score | mIoU  |
|-------------|-----------|--------|----------|-------|
| Baseline    | 75.5      | 28.4   | 61.3     | 53.0  |
| SGAN-SAL-SEED | 31.7      | 47.6   | 33.2     | 45.8  |
| SGAN-SEED   | 76.1      | 48.6   | 70.5     | 62.4  |
| SGAN\textsuperscript{CLS} | 74.4      | 60.0   | 72.0     | 63.7  |
| SGAN\textsuperscript{SEG} | 73.8      | 57.1   | 70.9     | 63.4  |
| SGAN        | 76.4      | 57.4   | 73.0     | 64.2  |

The F-measure is defined as the weighted harmonic mean of the precision and recall:

\[
F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}},
\]

where \(\beta^2\) is empirically set to be 0.4 to emphasize the importance of precision.

Table 5 reports the evaluation results. From it we get the following observations: (1) The SGAN-SAL-SEED model that applies the self-attention mechanism directly under weakly-supervised settings degrades the seed’s precision drastically and thus leading to poor segmentation performance. (2) The full model, SGAN, enhances the recall of seeds by a large margin while maintains the precision level, indicating that it can produce dense and accurate seeds. (3) The F-measure score shows a strong correlation with the final segmentation performance (mIoU). The higher a F-measure is, the better a variant SGAN model can perform.

3) VISUALIZATION OF CONTEXT ATTENTION MAPS

To better understand how the self-attention mechanism behaves in our models, Figure 5 visualizes the context attention maps learned by different variant of SGAN. Specifically, we select one discriminative pixel in each image and mark it by a yellow ‘%’. The attentions propagated from the selected pixel to all other pixels are indicated in the corresponding column of the learned context attention map. We reshape the column into the image size and overlay it on the color image for visualization. As shown in Figure 5, simply integrating the self-attention mechanism in a weakly-supervised network tends to mess up the attentions. Saliency priors can prohibit from spreading the attentions to the background. By further integrating the class-specific attention cues, our full model can restrict the attentions propagated mostly to the pixels belonging to the same category with the selected pixel. These maps help us to interpret the CAMs presented in Figure 4.

4) INFLUENCE OF THE WEIGHTING FACTOR \(\lambda\)

The weighting factor \(\lambda\) in the total loss (in Equation (7)) of SGAN determines the impact of the seed loss. Without the seed loss, no class-specific attention cues are included and our SGAN cannot handle the problem of mis-spreading attentions among foreground categories. Whereas, putting too much weight on this term may cause inefficient training due to the sparsity of the seeds. Therefore we carry out a set of experiments to check the influence of \(\lambda\) and report the results in Table 6. It shows that \(\lambda = 0.15\) leads to the best performance. Therefore, we empirically set this value throughout all other experiments.

5) INFLUENCE OF SALIENCY DETECTORS

As marked out in Table 1, it is quite common for WSSS methods to take saliency as additional guidance. The reason is that it can provide rough localization of foreground objects and therefore help the propagation of CAMs. In our work, we adopt it to prevent object’s discriminative information from spreading to unexpected background regions.

Various saliency detectors, such as DHSNet [45], DSS [46], and S-Net [29], have been adopted in recent WSSS methods [14], [18], [19] to produce saliency masks. In our work, saliency masks are generated by S-Net [29]. We also conduct experiments using DHSNet [45] and DSS [46]. Table 7 shows that SGAN is not so sensitive to saliency detectors, because these saliency detection methods perform comparable well on outdoor scenarios but all have difficulties in handling indoor scenes. In addition, we admit that the errors in saliency masks may have negative effects for the
V. CONCLUSION

In this paper, we have presented a saliency guided self-attention network to address the semantic segmentation problem supervised by image-level labels only. To generate dense and accurate object seeds, we introduced the self-attention mechanism into the weakly-supervised scenario and utilized both class-agnostic saliency maps and class-specific attention cues to enable the mechanism work effectively. Extensive experiments on PASCAL VOC 2012 and COCO show that the proposed method outperforms the baseline model with a large margin and performs better than all other state-of-the-art methods.

REFERENCES

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

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

[3] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “DenseASPP for semantic segmentation in street scenes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3684–3692.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241.

[5] Y. Yuan and J. Wang, “OCNet: Object context network for scene parsing,” 2018, arXiv:1809.00916. [Online]. Available: https://arxiv.org/abs/1809.00916

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019.

[7] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IJCV, 2019.

[8] J. Xu, A. G. Schwing, and R. Urtasun, “Learning to segment under various forms of weak supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3781–3790.

[9] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3519–3527.

[10] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in Proc. ECCV, Cham, Switzerland: Springer, 2016, pp. 695–711.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2921–2929.

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

[13] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, “Weakly-supervised semantic segmentation network with deep seeded region growing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7014–7023.
[38] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, “Augmented feedback in semantic segmentation under image level supervision,” in Proc. ECCV, 2016, pp. 90–105.

[39] A. Roy and S. Todorovic, “Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3529–3538.

[40] J. Ahn and S. Kwak, “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4981–4990.

[41] T. Zhang, G. Lin, J. Cai, T. Shen, C. Shen, and A. C. Kot, “Decoupled spatial neural attention for weakly supervised semantic segmentation,” IEEE Trans. Multimedia, vol. 21, no. 11, pp. 2930–2941, Nov. 2019.

[42] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, “Joint learning of saliency detection and weakly supervised semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2019.

[43] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, “Reliability does matter: An end-to-end weakly supervised semantic segmentation approach,” in Proc. AAAI, 2020.

[44] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, “Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1742–1750.

[45] N. Liu and J. Han, “DHSNet: Deep hierarchical saliency network for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 678–686.

[46] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, “Deeply supervised salient object detection with short connections,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 815–828, Apr. 2019.

QI YAO was born in Jiangxi, China, in 1995. He received the B.E. degree in information science and electronic engineering from Zhejiang University, Hangzhou, Zhejiang, in 2017, where he is currently pursuing the master’s degree in information and communication engineering. His main research interests include computer vision and deep learning.

XIAOJIN GONG (Member, IEEE) received the B.A. and M.A. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from Virginia Tech, USA, in 2009. She is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Her research interests include computer vision, image processing, and artificial intelligence.