Scale-Aware Feature Network for Weakly Supervised Semantic Segmentation

LIAN XU\(^1\), MOHAMMED BENNAMOUN\(^1\), (Senior Member, IEEE), FARID BOUSSAID\(^2\), AND FERDOUS SOHEL\(^3\), (Senior Member, IEEE)

\(^1\)Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia
\(^2\)Department of Electrical, Electronics and Computer Engineering, The University of Western Australia, Perth, WA 6009, Australia
\(^3\)College of Science, Health, Engineering and Education, Murdoch University, Perth, WA 6150, Australia

Corresponding author: Lian Xu (lian.xu@research.uwa.edu.au)

This work was supported in part by the China Scholarship Council and Australian Research Council under Grant DP150104251.

ABSTRACT Weakly supervised semantic segmentation with image-level labels is of great significance since it alleviates the dependency on dense annotations. However, as it relies on image classification networks that are only capable of producing sparse object localization maps, its performance is far behind that of fully supervised semantic segmentation models. Inspired by the successful use of multi-scale features for an improved performance in a wide range of visual tasks, we propose a Scale-Aware Feature Network (SAFN) for generating object localization maps. The proposed SAFN uses an attention module to learn the relative weights of multi-scale features in a modified fully convolutional network with dilated convolutions. This approach leads to efficient enlargements of the receptive fields of view and produces dense object localization maps. Our approach achieves mIoUs of 62.3\% and 66.5\% on the PASCAL VOC 2012 test set using VGG16 based and ResNet based segmentation models, respectively, outperforming other state-of-the-art methods for the weakly supervised semantic segmentation task.

INDEX TERMS Weakly supervised semantic segmentation, attention mechanism, object saliency, multi-scale learning.

I. INTRODUCTION

Human cognition is formed gradually through the process of the continuous exploration of our surroundings. A large body of research has focused on developing machines with the learning ability of humans. For semantic segmentation, the prediction accuracy relies heavily on a large number of pixel-level labels, which comes at a prohibitively high human annotation cost [1]. In contrast, humans perform semantic segmentation without the need of such fine pixel-level supervision and information. They only rely on weak supervision instead. Inspired by this observation, a number of weakly supervised semantic segmentation approaches have been proposed, using the following weak supervision categories: bounding boxes [2], [3], scribbles [4], [5], points [6] and image-level labels [7]–[10]. Among them, image-level labels are the most popular and cost-effective, as they are simple, and easy to collect.

Weakly supervised semantic segmentation involves two major tasks: the generation of pseudo or approximate segmentation ground-truth images and the training of the segmentation network. With image-level labels, prior works have relied on an image classification network to obtain object localization maps. Approaches such as Class Activation Map (CAM) [11] and the gradient based CAM (Grad-CAM) [12] can identify class-discriminative object regions from image classification networks. However, these methods generally produce coarse object localization maps and cannot discover the majority of the non-discriminative regions, especially for large-scale objects. That is because it is usually sufficient for the classification network to make correct classification prediction from small discriminative regions. However, the resulting object localization maps only provide limited supervision to the training of segmentation models, leading to inferior segmentation performance. Most works address this problem either by relying on additional networks [13], [14], or by requiring repetitive network training steps [15] to expand and refine original object localization maps. In contrast, this work aims to generate better object
localization maps by only relying on an image classification network.

In order to construct such a network, we propose an architecture based on a fully connected convolutional network (FCN) and dilated convolutions, which enlarge the receptive fields without decreasing the size of the feature maps [16]. The proposed architecture can thus provide rich context information to produce accurate object localization maps. Moreover, in order to address the problem of incomplete object localization maps (e.g., especially in the case of images with large-scale objects), we seek to extract useful information from multi-scale images, as inspired by [17], [18]. With such an approach, small-scale objects can be well localized in images of the original resolution, while large-scale objects will be associated to expanded localization maps when the original resolution of the input images is reduced. Therefore, we propose to adaptively aggregate feature maps from multi-scale input images to cope with objects of different scales. To this end, we propose a scale attention model that learns to assign a weight to each scale at each spatial location of the feature maps. The final object localization maps are then generated from the weighted sum of the feature maps across all scales.

The major contributions of this work can be summarized as follows:

- We introduce the concept of attention on scales and propose a Scale-Aware Feature Network (SAFN) to produce dense object localization maps for weakly supervised semantic segmentation.
- Our SAFN improves the completeness of the object localization maps by taking image scale into account. To the best of our knowledge, this is the first time that a multi-scale framework for weakly supervised semantic segmentation is proposed.
- Extensive experiments have been carried out to analyze the performance and demonstrate the effectiveness of the proposed SAFN. Our approach produces mIoUs of 65.4% and 66.5% on val and test sets of the PASCAL VOC 2012 segmentation benchmark, achieving the state-of-the-art performance.

The rest of this paper is organized as follows. In Section II, we explore the related works on CNN visualization, weakly supervised semantic segmentation and attention mechanisms. Section III describes the details of our method, which includes the network architecture of the proposed SAFN, the generation of object localization maps and pseudo segmentation ground-truth images. In Section IV, we report the experimental results of performance comparison with state-of-the-art techniques and ablation studies. Finally, a conclusion is drawn in Section V.

II. RELATED WORKS

Our work draws on recent CNN visualization works, weakly supervised learning and attention mechanisms.

A. CNN VISUALIZATION

Studies on CNN visualization or interpretability are of great importance because they aim to understand why CNNs have such a superior performance. Achievements in this area have opened up new promising research directions, such as weakly supervised object detection [19]–[22] and semantic segmentation.

“Deconvolution” [23] and “Guided Backpropagation” [24] are two classic visualization techniques using gradients to produce CNN saliency maps. These techniques differ in the way backpropagation occurs through the rectified linear unit (ReLU) [25]. Although the generated saliency maps show which part of inputs affect the prediction, they are not class-discriminative. Subsequently, a technique named class activation map (CAM) [11] was proposed to indicate the discriminative image regions that CNNs use to classify a specific class. CAM modifies the image classification CNN architecture by replacing all the fully connected layers except the last prediction layer with a Global Average Pooling (GAP) layer. The learned weights of the prediction layer reflect the importance of the last convolutional feature maps in the prediction of corresponding classes. The weighted feature maps are then summed up to form the class-specific object localization maps. Similar techniques based on Global Max Pooling (GMP) [26] and log-sum-exp pooling [27] have also been investigated. However, such approaches trade-off model complexity and performance for an improved understanding of the working of the model. To address this limitation, Grad-CAM [12] was proposed to provide visual explanations of deep networks without the need of modifying the architectures. This approach can highlight image regions that are important for prediction by using the gradients of the target on the last convolutional layers. A more generalized technique, “Grad-CAM++” [28], was proposed to address Grad-CAM’s limited ability to localize multiple occurrences of a class in an image. This is achieved by taking a weighted average of pixel-wise gradients when computing the weights for a particular class or an activation map. These visualization techniques have been the basis for a wide variety of downstream tasks, e.g., weakly supervised semantic segmentation.

B. WEAKLY SUPERVISED SEMANTIC SEGMENTATION

Approaches to weakly supervised semantic segmentation with image-level labels rely on image classification CNNs to produce initial object localization maps. These methods mainly aim to address the problem that the initial object localization maps are sparse and thus not dense enough for training a well-performing segmentation model.

One line of approaches discovers the non-discriminative regions based on common object features. Ahn [13] proposed “AffinityNet”, which takes an image as input and outputs the affinities of pairs of adjacent image pixels. The initial object localization maps can be expanded through random walk according to the generated affinity matrix. In [14], a region classification network is proposed to be trained on
superpixels that are labeled with the initial object localization cues. The trained network can be used to predict classes of unlabeled regions. In [29], based on the traditional seeded region growing method, regions are grown from the initial object localization cues depending on their proposed similarity criterion.

Another line of approaches expands object localization maps by improving CNN’s attention to non-discriminative image regions. The most prevalent method is the erasing strategy. Wei et al. [15] proposed an adversarial erasing method to progressively mine discriminative object regions. However, repeated training of the classification network is time-consuming and determining when to stop the erasing operation remains a problem. Li et al. [30] designed guided attention inference networks with two identical branches. One branch aims to discover the most discriminative regions from the original image. The other branch takes the original image minus the most discriminative regions that are discovered by the first branch, as input. An attention mining loss is proposed to constrain the probability of the erased image being correctly classified, thus forcing the other branch to discover the whole object of interest. Moreover, Wei et al. [9] took advantage of “dilated convolution” with an enlarged receptive field to incorporate context and achieve dense object localization maps, by adding multiple dilated convolutional blocks of different dilated rates to the image classification network. A recent trend is to use dropout [31] to zero-out features randomly. As a result, the network needs to look elsewhere for evidence to achieve the classification objective. DSNs [32] uses a two-branch module to decouple the input feature maps. One uses dropout layers to obtain expansive attention maps, fused with the normal discriminative attention maps from the other branch, to produce better object localization maps. FickleNet [33] uses a dropout layer to randomly select pairs of hidden units to predict the classification scores, thereby associating non-discriminative parts of an object with the discriminative parts of the same object.

Apart from methods improving object localization maps, several works focused on other means to improve weakly supervised semantic segmentation performance. For examples, Kolesnikov et al. [7] proposed a new loss function that brings up three principles including seed, expand and constrain (SEC) to guide the weakly supervised training of semantic segmentation models. Wei et al. [15] proposed an online PSL approach that uses class probabilities from an added classification branch as weights to produce weighted segmentation masks, which together with the pseudo segmentation ground-truth images serve for supervision. A recent work by Shimoda and Yanai [34] focuses on estimating the noises in segmentation results caused by post-processing methods such as conditional random fields, so as to improve the segmentation accuracy by removing estimated noises.

C. ATTENTION MECHANISM

Attention mechanism has primarily been used in a variety of NLP tasks [35]–[38] such as machine translation. Since an attention mechanism allows the network to manage its focus and improve the efficiency of the data processing, it has been widely applied in the computer vision field. An attention module embedded network is able to allocate attention unevenly across a field of information and focus on certain inputs while ignoring or diminishing the importance of others. In general, there are several methods of allocating importance for CNNs to achieve strong representational power in visual tasks. Hu et al. [39] won the top prize for the ILSVRC 2017 classification task for their proposed squeeze-and-excitation network (SENet). The SENet introduces a novel channel attention block to model the inter-dependencies between channels, which enables in turn an adaptive re-calibration of the channel-wise feature responses. In [40], both spatial and channel attention modules are used to boost the network performance of the classification and detection tasks. The residual attention network proposed in [41], shows the benefits of stacking attention modules with residual attention learning. In particular, attention on single point is shown to achieve better performance than spatial or channel attention.

III. PROPOSED SCALE-AWARE FEATURE NETWORK

We propose a scale-aware feature network to produce dense object localization maps. In this section, we start by introducing the proposed network architecture, and then elaborate the procedures for generating object localization maps and pseudo segmentation ground-truth images.

A. NETWORK ARCHITECTURE

We build our baseline model to generate object localization maps based on the VGG16 [42] network. The last two pooling layers are removed and the convolutional layers of the last block are replaced by the dilated convolutions with a dilated rate of 2 and a stride of 1. Furthermore, we replace the fully connected layers with the following cascaded layers: a convolutional layer with $3 \times 3$ filters and 1024 channels, a convolutional layer with $1 \times 1$ filters and 1024 channels and another convolutional layer with $1 \times 1$ filters and $N$ channels (where $N$ denotes the number of classes), followed by a global average pooling layer.

We propose to strengthen the baseline network by adding a scale-attention module for processing multi-scale features. Figure 1 illustrates the overall architecture of the proposed scale-aware feature network. Suppose the input image $I$ is scaled to $S$ scales $\{I^1, I^2, \ldots, I^S\}$, the output features from the penultimate layer (before the global average pooling layer) are $F^1, F^2, \ldots, F^S$, respectively. These feature maps are then re-sampled to be at the same size through bi-linear interpolation and stacked along the channel dimension. Finally, the stacked features are passed through an attention module. The attention module is composed of two convolutional layers: one with 512 $3 \times 3$ filters comes at first, followed by a second with $S 1 \times 1$ filters. The attention module produces $S$ 2D attention matrices $\{A^1, A^2, \ldots, A^S\}$. The scale
The attention module can learn to weigh the importance of features at each location and scale. It can thus effectively allocate attention to features at different locations and scales, achieving appropriate feature aggregation. Moreover, the attention module allows the gradients of the loss function to be back-propagated through. Therefore, it can be trained jointly with the multi-label image classification part in an end-to-end manner. The Multi-label Soft Margin Loss function (also referred to as Sigmoid Cross-entropy loss) is used as our multi-label image classification objective function:

\[ \mathcal{L}(x, y) = - \sum_{i=1}^{N} \left[ y_i \log \frac{\exp(x_i)}{1 + \exp(x_i)} + (1 - y_i) \log \frac{1}{1 + \exp(x_i)} \right], \tag{3} \]

where \( x \) is an \( N \)-dimensional vector, which is the output of the proposed network given an input image. \( N \) denotes the number of classes and \( y \) represents the \( k \)-hot encoding of multiple labels of the input image.

### B. GENERATING OBJECT LOCALIZATION MAPS

To generate object localization maps with image-level labels, we use a method which directly selects the localization map of each class from the last convolutional layers of the image classification network in a forward pass. This process is illustrated in Figure 2. Given that the feature maps before the global average pooling layer of the proposed SAFN are MSF, the object localization map that is relevant to class \( c \) is computed as:

\[ M_c = \frac{\text{ReLU}(\text{MSF}_c)}{\max(\text{ReLU}(\text{MSF}_c))}. \tag{4} \]

This approach is more convenient, compared to the most prevalent CAM-family techniques, which require an extra step, i.e., the multiplication of the feature maps by learned weights. This small module of generating object localization maps is easy to be embedded into any complex network. Similar methods are also adopted in [13], [43], [44] and theoretical details have been presented in [44] to show that this approach can provide the same-quality object localization maps as the CAM technique.

### C. PSEUDO SEGMENTATION GROUND-TRUTH GENERATION

The proposed SAFN is able to produce dense object localization maps but with ambiguous boundaries. Salient object detectors have been used to identify visually distinctive

---

**Algorithm 1 Pseudo Segmentation Ground-Truth Generation**

**Input:** C: Image labels; M: Object localization maps; S: Saliency maps; \( f \): combination function; \( \gamma \): Threshold.

**Output:** G: Pseudo segmentation ground-truth images.

1: for each \( c \) in C and each location \((i, j)\) do
2: \( H_c(i, j) \leftarrow f(M_c(i, j), S(i, j)) \)
3: end for
4: for each location \((i, j)\) do
5: if \( H_c(i, j) < \gamma \) then
6: \( G(i, j) \leftarrow 0 \) (0 denotes the background class)
7: else
8: \( G(i, j) \leftarrow \arg\max_c M_c(i, j) \)
9: end if
10: end for

---

**FIGURE 1.** Overall architecture of the proposed scale-aware feature network. GAP denotes global average pooling.
FIGURE 2. Illustration of the process of producing object localization maps with the proposed SAFN ($N$ is the number of classes).

FIGURE 3. Illustration of the proposed pseudo segmentation ground-truth generation.

objects from the background in an image [45]. In order to generate high-quality pseudo segmentation ground-truth images, we use a salient object detector to provide class-agnostic saliency maps for refining our object localization maps, as is done in most prior works [9], [15], [43], [46]. This procedure is illustrated in Figure 3. Specifically, we follow the combination function adopted in [43], but we do not use iterative erasing with the saliency detector to enhance its performance. As in [46], we use the off-the-shelf saliency object detector from [47] to generate saliency maps for training images. We compute the pixel-wise harmonic mean between object localization maps $M$ and class-agnostic saliency maps $S$ as:

$$H_c(i,j) = \frac{2}{\frac{1}{M_c(i,j)} + \frac{1}{S(i,j)}},$$

where $H_c(i,j)$ is the confidence score of class $c$ of the pixel at spatial location $(i,j)$. The hard thresholding is then applied on the class confidence scores to define the pseudo segmentation ground-truth. This procedure is summarized in Algorithm 1. In particular, object regions for which confidence scores are below a threshold $\gamma$ are regarded as background. Each of the rest pixels is assigned to the class with the largest confidence value. There are two types of conflicting pixels that we chose to ignore: (i) those regarded as background but with high saliency scores, as the saliency detector may discover object classes that are not needed in the semantic segmentation. (ii) those with more than one class confidence scores greater than a high value, indicating the case of potential boundaries between different classes.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

We evaluated our approach on the PASCAL VOC 2012 [48] which has been used as the benchmark dataset for state-of-the-art techniques [9], [13], [30], [33], [34], [43], [46]. The original dataset has 20 categories and one background category. It includes training, validation and testing splits with 1,464, 1,449 and 1456 images, respectively. Following common practice, we used the augmented dataset of 10,582 images with image-level annotations provided by [49] for training.

2) EVALUATION METRIC

The mean Intersection-over-Union (mIoU) of all 21 categories between the outputs from the segmentation network and pixel-level ground-truth is used to evaluate the performance on the val and test sets. The results on the test set are obtained from the official PASCAL VOC online evaluation server.

3) IMPLEMENTATION DETAILS

We used PyTorch [50] to implement our approach.1 For the multi-label image classification training part, the baseline network is pre-trained on ImageNet [51]. The entire network parameters are then fine-tuned on the PASCAL VOC 2012 dataset [48] using stochastic gradient descent (SGD). We optimized our proposed SAFN by minimizing a multi-label soft margin loss. For data augmentation, random horizontal flipping and color jittering were applied to input images, followed by random cropping to 448 × 448. We train the network for 15 epochs with a batch size of 4. The initial learning rate was set to 0.01 (0.1 for the last prediction layer and the attention module which were both trained from scratch), to which a polynomial decay was applied. For generating pseudo segmentation ground-truth images, we set the threshold $\gamma$ in Algorithm 1 as 0.2.

For the semantic segmentation part, to fairly compare with other works, we adopted VGG16 based Deeplab-LargeFOV [16] as our segmentation model, which was pre-trained on ImageNet. The training images were randomly cropped to $321 \times 321$ and horizontally flipped. We used a batch size of 20 and fine-tuned the segmentation model using our generated pseudo segmentation ground-truth images for 6K iterations. The initial learning rate was set as 0.001 and decreased by a factor of 10 every 2000 iterations. The SGD optimizer was used with a momentum of 0.9 and a weight decay of 0.0005. At test time, we used the maximum

1Codes will be released at https://github.com/lianxxx/SAFN once this paper is accepted.
TABLE 1. Performance comparison of weakly supervised semantic segmentation methods in terms of mIoU(%) on PASCAL VOC 2012 val and test sets. All methods listed below use VGG16 based segmentation models.

| Method                  | Training | val  | test |
|-------------------------|----------|------|------|
| Supervision: Image-level and additional annotations |          |      |      |
| MIL ($CVPR'15$) [27]    | 700K     | 42.0 | 40.6 |
| TransferNet ($CVPR'16$) [55] | 70K     | 52.1 | 51.2 |
| STC ($TPAMI'17$) [56]   | 50K      | 49.8 | 51.2 |
| CrailSeg ($CVPR'17$) [57] | 970K    | 58.1 | 58.7 |
| AISI ($ECCV'18$) [58]   | 30K      | 61.7 | 63.1 |

Supervision: Image-level annotations only

| Method                  | Training | val  | test |
|-------------------------|----------|------|------|
| SEC ($ECCV'16$) [7]     | 10K      | 50.7 | 51.7 |
| AE-PSL ($CVPR'17$) [15] | 10K      | 55.0 | 55.7 |
| Roy et al. ($CVPR'17$) [59] | 10K     | 52.8 | 53.7 |
| DCSP ($BMVC'17$) [43]   | 10K      | 58.6 | 59.2 |
| DSRG ($CVPR'18$) [29]   | 10K      | 59.0 | 60.4 |
| GAIN ($CVPR'18$) [30]   | 10K      | 55.3 | 56.8 |
| MCOF ($CVPR'18$) [14]   | 10K      | 56.2 | 57.6 |
| Ahn et al. ($CVPR'18$) [13] | 10K    | 58.4 | 60.5 |
| MDC ($CVPR'18$) [9]     | 10K      | 60.4 | 60.8 |
| SeeNet ($NIPS'18$) [46] | 10K      | 61.1 | 60.7 |
| FickleNet ($CVPR'19$) [33] | 10K    | 61.2 | 61.9 |
| SAFN (Ours)             | 10K      | 61.9 | 62.3 |

TABLE 2. Performance comparison of weakly supervised semantic segmentation methods in terms of mIoU(%) on PASCAL VOC 2012 val and test sets. All methods listed below use ResNet based segmentation models.

| Method                  | Training | val  | test |
|-------------------------|----------|------|------|
| Supervision: Image-level and additional annotations |          |      |      |
| Boot-Web ($CVPR'18$) [60] | 80K     | 63.0 | 63.9 |
| AISI ($ECCV'18$) [58]   | 30K      | 64.5 | 65.6 |

Supervision: Image-level annotations only

| Method                  | Training | val  | test |
|-------------------------|----------|------|------|
| DCSP ($BMVC'17$) [43]   | 10K      | 60.8 | 61.9 |
| DSRG ($CVPR'18$) [29]   | 10K      | 61.4 | 63.2 |
| MCOF ($CVPR'18$) [14]   | 10K      | 60.3 | 61.2 |
| Ahn et al. ($CVPR'18$) [13] | 10K    | 61.7 | 63.7 |
| SeeNet ($NIPS'18$) [46] | 10K      | 63.1 | 62.8 |
| FickleNet ($CVPR'19$) [33] | 10K    | 64.9 | 65.3 |
| SSD ($GCCV'19$) [34]    | 10K      | 64.9 | 65.5 |
| SAFN (Ours)             | 10K      | 65.4 | 66.5 |

values of results from multi-scale inputs (scales of 0.5, 0.75, 1.0, 1.25 and 1.5 were used in our experiments) as our final predictions. For post-processing, Conditional Random Fields (CRF) [52] with parameters (i.e., $b_i = 4$, $b_{xy\_std} = 67$, $b_{rgb\_std} = 3$, $pos_w = 3$, $pos_{xy\_std} = 1$), were also used to refine the predicted masks from the segmentation network. We also report results from ResNet [53] based Deeplab-ASPP [54] segmentation model.

B. COMPARISON WITH STATE-OF-THE-ART

Table 1 and 2 show comparisons with existing state-of-the-art methods in terms of mIoUs on PASCAL VOC 2012 val and test sets using VGG16 and ResNet based segmentation models, respectively. Our proposed SAFN achieves mIoUs of 61.9% and 62.3% on val and test sets using VGG16 based segmentation model, and achieves mIoUs of 65.4% and 66.5% using ResNet based segmentation model, both outperforming the state-of-the-art methods with the same-level annotations. Segmentation results of IoU per class are shown in Table 3.

There are some methods that use additional annotations or stronger supervisions than image-level labels, such as MIL [27] augments the image-level labeled training set by using around 700K class-relevant images from ImageNet, while TransferNet [55] uses around 70K class-irrelevant but pixel-level labeled images from the MS-COCO dataset. AISI [58] uses an instance-level salient object detector to help generate pseudo segmentation ground-truth images, while others commonly use saliency maps that only differentiate foreground from background without instance information. Webly images and videos have been used as a source of training data of weakly supervised learning as they are available in massive quantity and easy to acquire through keyword search, such as in STC [56], CrailSeg [57] and Boot-Web [60]. Among these methods, the use of the instance-level salient object detector in AISI [58] shows its superiority in boosting the performance. In comparison, despite using the common saliency maps and only 1/3 of its training images, our method achieves comparable and better performance than AISI [58] when using VGG16 and ResNet based segmentation models, respectively.

Examples of qualitative results from the proposed SAFN are shown in Figure 4. It can be observed that trained with our generated pseudo ground-truth images, the model can produce good segmentation results for images of various scenes with multiple objects. In comparison, ResNet-based model shows its superiority in segmenting objects with large inter-class similarities, such as the cow and the horse in the second and third rows. There are some failure cases shown in the last row where the table and some chairs were mis-classified as background.

C. ABLATION STUDIES

1) COMPARISON OF DIFFERENT AGGREGATION METHODS

To evaluate the benefits of the attention module in our proposed SAFN, we additionally implement two different methods, i.e., max-pooling and average-pooling, for feature aggregation from multi-scale inputs (three scales, i.e., $\{1, 0.75, 0.5\}$, are used in the following experiments). Table 4 shows the comparison of mIoU scores on the PASCAL VOC val set from using different aggregation methods to generate object localization maps. It can be observed that our attention module produces better segmentation results for images of various scenes with multiple objects. In comparison, ResNet-based model shows its superiority in segmenting objects with large inter-class similarities, such as the cow and the horse in the second and third rows. There are some failure cases shown in the last row where the table and some chairs were mis-classified as background.
FIGURE 4. Qualitative results of the proposed SAFN on the PASCAL VOC 2012 val set. (a) Input images. (b) Ground-truth images. (c) Our results from VGG16-based segmentation model. (d) Our results from ResNet-based segmentation model.

TABLE 3. Performance of the proposed SAFN in terms of IoUs (%) on PASCAL VOC 2012 val and test sets using VGG16 and ResNet based segmentation models.

|       | bgk | plane | bike | bird | boat | bottle | bus | cat | chair | cow | table | dog | horse | milk | person | plant | sheep | sofa | train | p | mIoU |
|-------|-----|-------|------|------|------|--------|-----|-----|-------|-----|-------|-----|-------|------|--------|-------|-------|------|-------|---|-----|
| VGG16-val | 90.0 | 83.7 | 32.8 | 75.2 | 58.4 | 64.8 | 80.1 | 74.0 | 82.2 | 15.1 | 69.9 | 21.9 | 76.0 | 70.0 | 69.6 | 77.3 | 36.2 | 26.4 | 74.1 | 26.8 | 67.0 | 53.9 | 61.9 |
| VGG16-test | 90.4 | 76.6 | 35.2 | 77.3 | 49.8 | 59.8 | 77.1 | 73.0 | 81.8 | 20.3 | 64.6 | 30.3 | 74.5 | 69.6 | 78.6 | 76.7 | 46.2 | 76.6 | 28.0 | 67.2 | 54.0 | 62.3 |
| ResNet-val | 90.6 | 84.5 | 37.3 | 81.2 | 63.3 | 70.7 | 78.7 | 72.6 | 87.6 | 20.5 | 76.4 | 15.7 | 79.7 | 76.4 | 74.6 | 79.0 | 46.8 | 79.1 | 28.6 | 72.0 | 59.1 | 65.4 |
| ResNet-test | 91.1 | 84.0 | 37.1 | 84.2 | 53.4 | 71.0 | 77.4 | 78.0 | 84.5 | 25.3 | 75.7 | 22.8 | 81.6 | 82.9 | 80.2 | 80.5 | 45.7 | 83.1 | 27.9 | 75.2 | 55.2 | 66.5 |

TABLE 4. Evaluation of different aggregation methods to generate pseudo ground truth images in terms of segmentation mIoUs (%) on the PASCAL VOC2012 val set. Results were obtained without CRF.

| Aggregation Methods | Segmentation inputs | Single scale | Multiple scales |
|---------------------|---------------------|--------------|-----------------|
| Max-pooling         | 57.2                | 59.3         |                 |
| Average-pooling     | 57.4                | 59.2         |                 |
| Attention (Ours)    | 57.6                | 59.5         |                 |

Figure 5 shows examples of small-scale objects (rows (1)-(3)) and large-scale objects (rows (4)-(6)), their object localization maps generated by using different input images that are resized to different scales, and the multi-scale aggregated results produced by different methods. As shown in column (b), small-scale objects can be well localized in the original image (scale = 1.0), while only small dis-
L. Xu et al.: Scale-Aware Feature Network for Weakly Supervised Semantic Segmentation

FIGURE 5. Visualization of object localization maps of input images with objects of different scales. (a) Input images. (b)-(d) Object localization maps generated by using different input images that are resized to different scales (scale = 0.5, 0.75, 1.0). (e)-(g) Multi-scale aggregated object localization maps produced by max-pooling, average-pooling and our proposed SAFN. (h) Ground truth.

Table 6. Comparison of the proposed SAFN with different scales in terms of segmentation mIoUs (%) on the PASCAL VOC2012 val set. Results were obtained without CRF.

| Segmentation inputs     | Single scale | Multiple scales |
|-------------------------|--------------|-----------------|
| Baseline (s = 1)        | 56.8         | 58.5            |
| SAFN (s = 1,0.75)       | 57.2         | 59.2            |
| SAFN (s = 1,0.75,0.5)   | **57.6**     | **59.5**        |

2) COMPARISON OF RESULTS FROM SAFN WITH DIFFERENT SCALES

Since small-scale objects can easily be completely localized while large-scale objects are often partially localized, we use additional scales that are between 0 and 1 to down-sample original images to form multi-scale inputs. Considering images of very small scales contain limited useful information, we use scales of 0.75 and 0.5 in our experiments. Table 6 reports performance comparisons between our baseline approach using single-scale images and the proposed SAFN using images of multiple scales. The proposed SAFN outperforms the baseline method by a reasonable margin under both segmentation settings. That indicates the benefits of our proposed multi-scale approach.

3) VISUAL COMPARISON OF OBJECT LOCALIZATION MAPS

Figure 6 shows a visual comparison of object localization maps produced by VGG16 using Grad-CAM, our baseline network, and our proposed SAFN. It can be observed that compared to VGG16, our methods infer more object regions and with more precise locations. For example, additional parts of the large-scale objects are detected, as shown in the first four rows. Also, in the fifth input image the second person at the back is undetected by VGG16, but it has been
detected with our methods. For the last image with the single label of “bottle”, VGG16 fails to localize the bottle, our baseline detects the bottle but with some other background objects, while our proposed SAFN locates it correctly. This demonstrates the effectiveness of the modified architecture of our baseline method, and the advantages of the proposed multi-scale representations of the proposed SAFN.

V. CONCLUSION
In this work, we propose a scale-aware feature network to produce object localization maps with image-level labels for the task of weakly supervised semantic segmentation. Our method takes advantage of multi-scale features from multi-scale images and adaptively aggregates multi-scale features through an attention mechanism, providing complete object localization maps. Experimental results demonstrate that the proposed SAFN achieves the state-of-the-art performance on both val and test sets of PASCAL VOC 2012 benchmark.

ACKNOWLEDGMENT
The authors acknowledge NVIDIA for providing a Titan X GPU for the experiments involved in this research.

REFERENCES
[1] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, “A guide to convolutional neural networks for computer vision,” Synth. Lectures Comput. Vis., vol. 8, no. 1, pp. 1–207, Feb. 2018.
[2] J. Dai, K. He, and J. Sun, “BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1635–1643.
[3] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, “Simple does it: Weakly supervised instance and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 876–885.
[4] D. Lin, J. Dai, H. Jia, S. He, and J. Sun, “ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3159–3167.
[5] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, “Normalized cut loss for weakly-supervised CNN segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1818–1827.
[6] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, “What’s the point: Semantic segmentation with point supervision,” in Proc. ECCV. Cham, Switzerland: Springer, 2016, pp. 549–565.
M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for...
J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving...
H. Cholakkal, J. Johnson, and D. Rajan, “Backtracking spatial pyra-
Y. Wei, Z. Shen, B. Cheng, H. Shi, Z. Jie, J. Feng, and T.S. Huang, “Revisiting dilated...
R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and T. Huang, “TS2C: Tight box mining with surrounding segmentation...
L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected Crfs,” in ICLR, Dec. 2015.
A. Mahmood, M. Bensamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. B. Fisher, “Coral classification with hybrid feature representations,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2016, pp. 519–523.
L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, “Attention to scale: Scale-aware semantic image segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3640–3649.
Q. Tao, H. Yang, and J. Cai, “Exploiting Web images for weakly supervised object detection,” IEEE Trans. Multimedia, vol. 21, no. 5, pp. 4115–4126, May 2019.
Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang, “TSSC: Tight box mining with surrounding segmentation context for weakly supervised object detection,” in Proc. ECCV, Sep. 2018, pp. 434–450.
H. Cholakkal, J. Johnson, and D. Rajan, “Backtracking spatial pyramid pooling-based image classifier for weakly supervised Top-Down salient object detection,” IEEE Trans. Image Process., vol. 27, no. 12, pp. 6064–6078, Dec. 2018.
K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, “Weakly supervised salient object detection by learning a classifier-driven map generator,” IEEE Trans. Image Process., vol. 28, no. 11, pp. 5435–5449, Nov. 2019.
M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Proc. ECCV. Cham, Switzerland: Springer, 2014, pp. 818–831.
J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in Proc. ICLR Workshop Track, Dec. 2014.
A. Mahendran and A. Vedaldi, “Salient deconvolutional networks,” in Proc. ECCV. Cham, Switzerland: Springer, 2016, pp. 120–135.
M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free?—Weakly-supervised learning with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 685–694.
P. O. Pinheiro and R. Collobert, “From image-level to pixel-level labeling with convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1713–1721.
A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized Gradient-Based visual explanations for deep convolutional networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 839–847.
[54] 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.

[55] S. Hong, J. Oh, H. Lee, and B. Han, “Learning transferable knowledge for semantic segmentation with deep convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3204–3212.

[56] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, “STC: A simple to complex framework for weakly-supervised semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[57] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, “Weakly supervised semantic segmentation using Web-crawled videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7322–7330.

[58] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, “Associating inter-image salient instances for weakly supervised semantic segmentation,” in Proc. ECCV, Sep. 2018, pp. 367–383.

[59] 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.

[60] T. Shen, G. Lin, C. Shen, and I. Reid, “Bootstrapping the performance of webly supervised semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1363–1371.

LIAO YU received the M.S. degree from Hainan University, China. She is currently pursuing the Ph.D. degree with the University of Western Australia. Her research interests include computer vision and deep learning.

MOHAMMED BENNAMOUN (Senior Member, IEEE) is Winthrop Professor with the Department of Computer Science and Software Engineering, UWA and is currently a Researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published four books (available on Amazon, one edited book, one Encyclopedia article, 14 book chapters, 120+ journal articles, 250+ conference publications, 16 invited and keynote publications. His h-index is 54 and his number of citations is 12,500+ (Google Scholar). He was awarded 65+ competitive research grants, from the Australian Research Council, and numerous other Government, UWA and industry Research Grants. He successfully supervised +26 the Ph.D. students to completion. He won the Best Supervisor of the Year Award at QUT, in 1998, and received award for research supervision at UWA, in 2008 and 2016, and Vice-Chancellor Award for mentorship, in 2016. He delivered conference tutorials at major conferences, including: IEEE COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2016), Interspeech 2014, the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) and European Conference on Computer Vision (ECCV). He was also invited to give a Tutorial at an International Summer School on Deep Learning (DeepLearn 2017).

FARID BOUSSAID received the M.S. and Ph.D. degrees in microelectronics from the National Institute of Applied Science (INSA), Toulouse, France, in 1996 and 1999, respectively. He joined Edith Cowan University, Perth, Australia, as a Postdoctoral Research Fellow, and a Member of the Visual Information Processing Research Group, in 2000. He joined the University of Western Australia, Crawley, Australia, in 2005, where he is currently a Professor. His current research interests include neuromorphic engineering, smart sensors, and machine learning.

FERDOUS SOHEL (Senior Member, IEEE) received the Ph.D. degree from Monash University, Australia, in 2009. He is currently an Associate Professor in information technology with Murdoch University, Australia. Prior to his joining Murdoch University, he was a Research Assistant Professor/Research Fellow with the University of Western Australia, from 2008 to 2015. His research interests include computer vision, image processing, pattern recognition, multimodal biometrics, scene understanding, robotics, and video coding. He was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2020–2021). He is a Member of Australian Computer Society.

* * *