Guided Attention Inference Network

Kunpeng Li, *Student Member, IEEE*, Ziyan Wu, *Member, IEEE*, Kuan-Chuan Peng, *Member, IEEE*, Jan Ernst, *Member, IEEE*, and Yun Fu, *Fellow, IEEE*

Abstract—With only coarse labels, weakly supervised learning typically uses top-down attention maps generated by back-propagating gradients as priors for tasks such as object localization and semantic segmentation. While these attention maps are intuitive and informative explanations of deep neural network, there is no effective mechanism to manipulate the network attention during learning process. In this paper, we address three shortcomings of previous approaches in modeling such attention maps in one common framework. First, we make attention maps a natural and explicit component in the training pipeline such that they are end-to-end trainable. Moreover, we provide self-guidance directly on these maps by exploring supervision from the network itself to improve them towards specific target tasks. Lastly, we proposed a design to seamlessly bridge the gap between using weak and extra supervision if available. Despite its simplicity, experiments on the semantic segmentation task demonstrate the effectiveness of our methods. Besides, the proposed framework provides a way not only explaining the focus of the learner but also feeding back with direct guidance towards specific tasks. Under mild assumptions our method can also be understood as a plug-in to existing convolutional neural networks to improve their generalization performance.

Index Terms—Convolutional neural network, semantic segmentation, network attention, weakly supervised learning, biased data

1 Introduction

Weakly supervised learning [2], [6], [12], [19], [23], [32], [41], [49], [50], [51], [52], [53], [54], [59] has recently become a popular research direction since it directly addresses the labeled data scarcity issue in computer vision. For instance, using only image-level labels, one can obtain the attention maps for a given input image with back-propagation on a Convolutional Neural Network (CNN). These maps are highly related to the network's response given specific patterns and tasks it was trained for. The intensity of each pixel on the attention maps indicates the degree that the corresponding pixel in the input image supporting the network's final output. From such attention maps without the need of pixel-level labels, it is already known that one can obtain the information of segmentation and localization [12], [59].

Although the attention maps are supervised by only classification loss without pixel-level labels, the attention maps generated by the trained network usually cover only the small and discriminative areas of the object of interest [22], [43], [59]. These attention maps can still provide useful localization cues as the priors for tasks like segmentation [23]. However, we believe that encouraging the attention maps to cover the target foreground objects as complete as possible can further improve the performance. Aligned with our belief, existing works either depend on consolidating attention maps from multiple networks [22] or aggregating multiple attention maps from a network via iterative erasing approaches [47]. Instead of post-processing the attention of the trained network passively, we propose an end-to-end framework where the attention of the network is trained jointly with the task-specific supervision towards improving the performance of the objective of the target task.

Moreover, as an effective tool to explain the network's decisions, attention maps can also help to find the biases of the trained network. For example, in the classification task where only image-level labels are available, it is likely that we encounter a pathological bias in the training data when the foreground object incidentally always correlates with the same background object (also mentioned in [39]). Fig. 1 shows an example image belonging to the class "boat" where it is highly likely that water and boats coexist. In this scenario there is no incentive in training to focus the network's attention on only the boat because focusing on water can also result in reasonably good performance. However, the generalization performance may suffer when the testing data does not maintain the coexistence relationship of foreground and background objects (e.g., boats out of water). Although such bias can be manually alleviated by re-balancing the training data, we propose to make the attention map trainable in our framework. As one benefit of this we are able to control the attention explicitly and can put manual effort in providing minimal supervision of attention rather than manually re-balancing the dataset. While it may not always be clear how to manually balance datasets to avoid bias, it is usually straightforward to guide attention to the regions of interest, e.g., using a small amount of pixel-level annotations like segmentation masks or bounding boxes. We also observe that our explicit self-guided attention model already improves the generalization performance even without extra supervision.

- *K. Li and Y. Fu are with Northeastern University, Boston, MA 02115. E-mail: {kunpengli, yunfu}@ece.neu.edu.*
- *Z. Wu, K.-C. Peng, and J. Ernst are with Siemens Corporate Technology, Princeton, NJ 08540. E-mail: wuzy.buaa@gmail.com, {kuanchuan.peng, jan.ernst}@siemens.com.*

Manuscript received 17 Oct. 2018; revised 26 Mar. 2019; accepted 22 May 2019. Date of publication 7 June 2019; date of current version 3 Nov. 2020. (Corresponding author: Kunpeng Li.)
Recommended for acceptance by S. Wang.
Digital Object Identifier no. 10.1109/TPAMI.2019.2921543

0162-8828 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

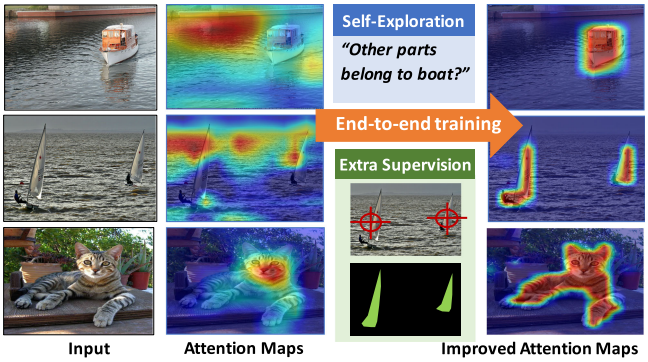


Fig. 1. The proposed Guided Attention Inference Network (GAIN) makes the network's attention on-line trainable and can plug in different kinds of supervision directly on attention maps in an end-to-end way. We explore the self-guided supervision from the network itself and propose GAIN $^{p}_{ext}$ when extra supervision are available. These guidance can optimize attention maps towards the task of interest.

This paper is an extended version of our previous work [27]. In particular, (a) we propose an extended framework for Guided Attention Inference Network (GAIN) so that it can be integrated with available bounding box annotations. (b) We use a new weakly-supervised semantic segmentation framework which takes the improved attention maps generated by our GAIN frameworks as input and achieves better performance on PASCAL VOC 2012 benchmark compared to our previous methods and other state-of-the-art algorithms. (c) We include more real-world experiments to demonstrate the effectiveness of GAIN in improving model generalization. (d) We include additional detailed analysis, more quantitative and qualitative results to help comprehensively analyze the performance of the proposed GAIN frameworks and better understand how it can be applied and extended to other applications.

According to our experimental results in the semantic segmentation task, our method achieves mIoU 59.4 and 59.6 percent, respectively on the PASCAL VOC 2012 segmentation *test* and *val* sets. Our method also outperforms the comparable state-of-the-art when limited pixel-level supervision (mIoU of 64.1 and 64.4 percent respectively) or bounding box supervision (mIoU of 62.6 and 62.8 percent respectively) is available during training.

The rest of our paper is organized as follows. Section 2 introduces related work in network attention modeling and weakly supervised learning. Section 3 describes the design of our Guided Attention Inference Network and its extensions to seamlessly integrate weak labels with stronger supervisions such as bounding boxes or pixel-level segmentation masks. Section 4 presents the experimental results of semantic segmentation task on VOC 2012 dataset. Section 5 describes the experiments we conducted to demonstrate how the propose GAIN pipeline can improve model generalization.

2 Related Work

Deep convolutional neural networks (DCNNs) have achieved great success in many areas recently [25], [26], [55], [56], [57]. Instead of just treating them as black boxes, various methods have been proposed to explain and analyze how DNN works from different views [6], [41], [51]. Visual attention is one of the efficient way which can explain the network's decision by highlighting the regions of images that are responsible for it. In this section, we discuss the most relevant work on network attention analysis, usage of attention maps in weakly-supervised methods especially for weakly-supervised semantic segmentation and how network attention can help to deal with dataset bias.

Network Attention. Visualization of model attention in Convolutional Neural Networks (CNNs) has been explored for network reasoning of visual recognition. Gradient back propagation based methods [41], [44], [51] interprets the gradient of the prediction score of a particular class respecting to the original input image. They visualize the network attention by locate regions that are helpful for predicting a class in such a way. [6] further proposes a feedback method to capture the top-down neural attention. According to the high-level semantic labels of the input image, [6] uses a feedback loop in the form of binary nodes between layers is introduced to infer the activation status of hidden layer neurons. CAM [59] adds an average pooling layer (GAP) after the last convolution layer of a CNN and applies a weighted sum of the last convolutional feature maps to obtain the attention maps. Excitation Backprop [52] is proposed based on a top-down visual attention model for human. As a novel back propagation method, it can pass along signals from top to down in the network hierarchy to locate the network attention for any CNN architecture using nonlinearities producing non-negative activations. The Excitation Backprop method is also extended to explain Recurrent Neural Network (RNN)-based models [4] to handle more complex recurrent, spatio-temporal dependencies. More recently, in order to deal with the drawback of CAM [52] that needs to change the structure of a CNN, Grad-CAM [39] extends it to many different available CNN architectures for tasks like image captioning and Visual Question Answering (VQA). Grad-CAM++ is then proposed to further improve Grad-CAM in terms of explaining occurrences of multiple objects in one single image as well as better object localization. Different from all these existing works that are trying to find a reasonable way to explain the network decision, we propose an end-to-end model to provide supervision on the network learning through these explanations, specifically attention mechanism. We validate that our method can guide the network to focus on the regions we expect without changing the network structure or learning extra parameters. The proposed guided attention learning will then benefit the corresponding visual task.

Network Attention for Weakly-Supervised Methods. Manually producing segmentation masks or bounding box annotations is a time-consuming task [5], [23]. To solve this issue, a lot of previous research studies how to train segmentation or detection models from weaker forms of annotation such as image-level labels, as this form of weak supervision can be collected very efficiently [23]. Recent works show that learning from only image-level labels, attention maps of a trained classification network can provide localization information for weakly-supervised object localization [32], [53], [54], [59], object detection [12], [49], semantic segmentation [2], [19], [23], [50], instance segmentation [60] etc. However, only trained with classification loss, the attention map only

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

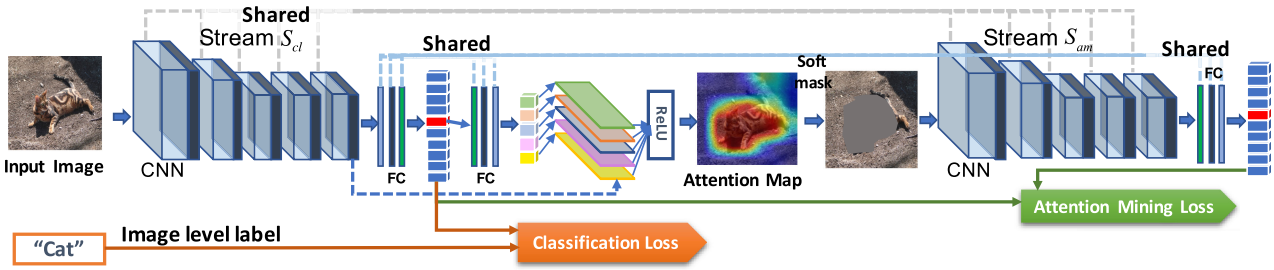


Fig. 2. GAIN has two streams of networks, S_{cl} and S_{am}, sharing parameters. S_{am} encourages the attention map to include regions contributing to the classification decision as complete as possible. The attention map is on-line generated and optimized by the two loss functions jointly.

covers small and most discriminative regions of the object of interest, which deviates from the requirement of these tasks that needs to localize interior, dense and complete regions. To mitigate this gap, [43] proposes a data augment method by randomly hiding regions in the training image. Then the network will be forced to seek other relevant parts when the most discriminative part is hidden. However, it relies on a strong assumption about the size of foreground objects (i.e., the size of the whole object should be bigger than that of patches). [47] proposes an adversarial erasing (AE) method to repetitive erasing the most discriminative regions of the input image and discover the rest part of object. Dense object regions are finally obtained by combining attention maps of AE step. Similarly, [22] trains two networks that focus on different parts of the object by using a two-phase learning strategy. First, a fully convolutional network (FCN) [31] is trained to locate the most discriminative regions of the object in an image. These most discriminative regions are then used to hide parts of feature maps of the second FCN, which forces it to focus on the rest important regions of the object. However, these methods require combination of attention maps from multiple classification networks. Attention maps of a single network still only focus on the most discriminative region. More recently, [50] revises a standard classification network by adding multiple dilated convolutional blocks of different dilation rates. Varying dilation rates can effectively transfer the surrounding discriminative information to non-discriminative object regions, which helps to achieve dense object localization using only one classification network. Fundamentally different from these approaches, the proposed method GAIN can provide supervision directly on network's attention in an end-to-end way. According to different levels of available supervision, we design corresponding loss functions to guide the network to focus on complete regions of interest. Therefore, using our methods, the attention maps of a single network are already more complete and improved without needs to change the network structure.

Attention Mechanism for Biased Data. Analyzing network attention can also help to identify bias in datasets. [39], [46] propose a way to find out the dataset bias by analyzing the location of attention maps of a trained model, which helps them to remove these bias by adding new samples to the dataset. However, in practical applications, it is time-consuming to build a new dataset and sometimes it is even hard to obtain samples that can remove all bias. How to guarantee the generalization ability of the learned network is still challenging. Our model can provide supervision directly on network's attention and guiding the network to focus on the areas critical to the task of interest. Therefore, our trained model is more robust to the dataset bias.

3 Proposed Method—GAIN

Since attention maps reflect the areas on input image which support the network's prediction, we propose the guided attention inference networks, which aims at supervising attention maps when we train the network for the task of interest. In this way, the network's prediction is based on the areas which we expect the network to focus on. We achieve this by making the network's attention trainable in an end-to-end fashion, which hasn't been considered by any other existing works [22], [39], [43], [47], [52], [59]. In this section, we describe the design of GAIN and its extensions tailored towards tasks of interest.

3.1 Self-Guidance on the Network Attention

As mentioned in Section 1, attention maps of a trained classification network can be used as priors for weakly-supervised semantic segmentation methods. However, purely supervised by the classification loss, attention maps usually only cover small and most discriminative regions of object of interest. These attention maps can serve as reliable priors for segmentation but a more complete attention map can certainly help improving the overall performance.

To solve this issue, our GAIN builds constraints directly on the attention map in a regularized bootstrapping fashion. As shown in Fig. 2, GAIN has two streams of networks, classification stream S_{cl} and attention mining S_{am}, which share parameters with each other. The constrain from stream S_{cl} aims to find out regions that help to recognize classes. The stream S_{am} is making sure that all regions which contribute to the classification decision will be included in the network's attention. In this way, attention maps become more complete, accurate and tailored for the segmentation task. The key here is that we make the attention map on-line generated and trainable by the two loss functions jointly.

Based on the fundamental framework of Grad-CAM [39], we streamlined the generation of attention map. An attention map corresponding to the input sample can be obtained within each inference so it becomes trainable in training stage. In stream S_{cl}, for a given image I, let $f_{l,k}$ be the activation of unit k in the lth layer. For each class c from the ground-truth label, we compute the gradient of score s^c corresponding to class c, with respect to activation maps of $f_{l,k}$. These gradients flowing back will pass through a global

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

average pooling layer [28] to obtain the neuron importance weights $w^c_{l,k}$ as defined in Eq. (1).

$$w^c_{l,k} = GAP\left(\frac{\partial s^c}{\partial f_{l,k}}\right), \tag{1}$$

where $GAP(\cdot)$ means global average pooling operation.

Here, we do not update parameters of the network after obtaining the $w^c_{l,k}$ by back-propagation. Since $w^c_{l,k}$ represents the importance of activation map $f_{l,k}$ supporting the prediction of class c, we then use weights matrix w^c as the kernel and apply 2D convolution over activation maps matrix f_l in order to integrate all activation maps, followed by a ReLU operation to get the attention map A^c with Eq. (2). The attention map is now on-line trainable and constrains on A^c will influence the network's learning

$$A^c = \text{ReLU}\ (\text{conv}(f_l, w^c)), \tag{2}$$

where l is the representation from the last convolutional layer whose features have a good balance between detailed spatial information and high-level semantics [41].

We then use the trainable attention map A^c to generate a soft mask to be applied on the original input image, obtaining I^{*c} using Eq. (3). I^{*c} represents the regions beyond the network's current attention for class c.

$$I^{*c} = I - (T(A^c) \odot I), \tag{3}$$

where $\odot$ denotes element-wise multiplication. $T(A^c)$ is a masking function based on a thresholding operation. In order to make it derivable, we use Sigmoid function as an approximation defined in Eq. (4).

$$T(A^c) = \frac{1}{1 + \exp(-\omega(A^c - \boldsymbol{\sigma}))}, \tag{4}$$

where $\boldsymbol{\sigma}$ is the threshold matrix whose elements all equal to σ. ω is the scale parameter ensuring $T(A^c)_{i,j}$ approximately equals to 1 when ${A^c}_{i,j}$ is larger than σ, or to 0 otherwise.

I^{*c} is then used as input of stream S_{am} to obtain the class prediction score. Since our goal is to guide the network to focus on all parts of the class of interest, we are enforcing I^{*c} to contain as little feature belonging to the target class as possible, i.e., regions beyond the high-responding area on attention map area should include ideally not a single pixel that can trigger the network to recognize the object of class c. From the loss function perspective it is trying to minimize the prediction score of I^{*c} for class c. To achieve this, we design the loss function called Attention Mining Loss as in Eq. (5).

$$L_{am} = \frac{1}{n}\sum_c s^c(I^{*c}), \tag{5}$$

where $s^c(I^{*c})$ denotes the prediction score of I^{*c} for class c. n is the number of ground-truth class labels for this image I.

As defined in Eq. (6), our final self-guidance loss L_{self} is the summation of the classification loss L_{cl} and L_{am}.

$$L_{self} = L_{cl} + \alpha L_{am}, \tag{6}$$

where L_{cl} is for multi-label and multi-class classification and we use a multi-label soft margin loss here. Alternative loss functions can be use for specific tasks. α is the weighting parameter. We use $\alpha = 1$ in all of our experiments.

With the guidance of L_{self}, the network learn to extend the focus area on input image contributing to the recognition of target class as much as possible, such that attention maps are tailored towards the task of interest, i.e., semantic segmentation. The joint optimization also prevents to erase all pixels. We verify the efficacy of GAIN with self guidance in Section 4.

3.2 GAIN$^p_{ext}$: Integrating Extra Pixel-Level Supervision

In addition to letting networks explore the guidance of the attention map by itself, we can also tell networks which part in the image they should focus on by using a small amount of extra supervision to control the attention map learning process. Based on this idea of imposing additional supervision on attention maps, we introduce the extension of GAIN: GAIN$^p_{ext}$, which can seamlessly integrate extra supervision in our weakly supervised learning framework.

Following Section 3.1, we still use the weakly supervised semantic segmentation task as an example application to explain the GAIN$^p_{ext}$. The way to generate trainable attention maps in GAIN$^p_{ext}$ during training stage is the same as that in the self-guided GAIN. In addition to L_{cl} and L_{am}, we design Attention Loss L_p based on the given external supervision. We define L_p as

$$L_p = \frac{1}{n}\sum_c (A^c - H^c)^2, \tag{7}$$

where H^c denotes the extra supervision, e.g., pixel-level segmentation masks in our example case.

Since generating pixel-level segmentation maps is extremely time consuming, we are more interested in finding out the benefits of using only a very small amount of data with external supervision, which fits perfectly with the GAIN$^p_{ext}$ framework shown in Fig. 3, where we add an external stream S^p_e, and these three streams share all parameters. Input images of stream S^p_e include both image-level labels and pixel-level segmentation masks. One can use only a very small amount of pixel-level labels through stream S^p_e to already gain performance improvement with GAIN$^p_{ext}$ (in our experiments with GAIN$^p_{ext}$, only $1\sim 10$ percent of the total labels used in training are pixel-level labels). The input of the stream S_{cl} includes all images in the training set with only image-level labels.

The final loss function, L_{ext-p}, of GAIN$^p_{ext}$ is defined as follows:

$$L_{ext-p} = L_{cl} + \alpha L_{am} + \omega L_p, \tag{8}$$

where L_{cl} and L_{am} are defined in Section 3.1, and ω is the weighting parameter depending on how much emphasis we want to place on the extra supervision (we use $\omega = 10$ in our experiments).

GAIN$^p_{ext}$ can also be easily modified to fit other tasks. Once we get activation maps $f_{l,k}$ corresponding to the network's final output, we can use L_p to guide the network to focus on areas critical to the task of interest. In Section 5, we show an example of such modification to guide the

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

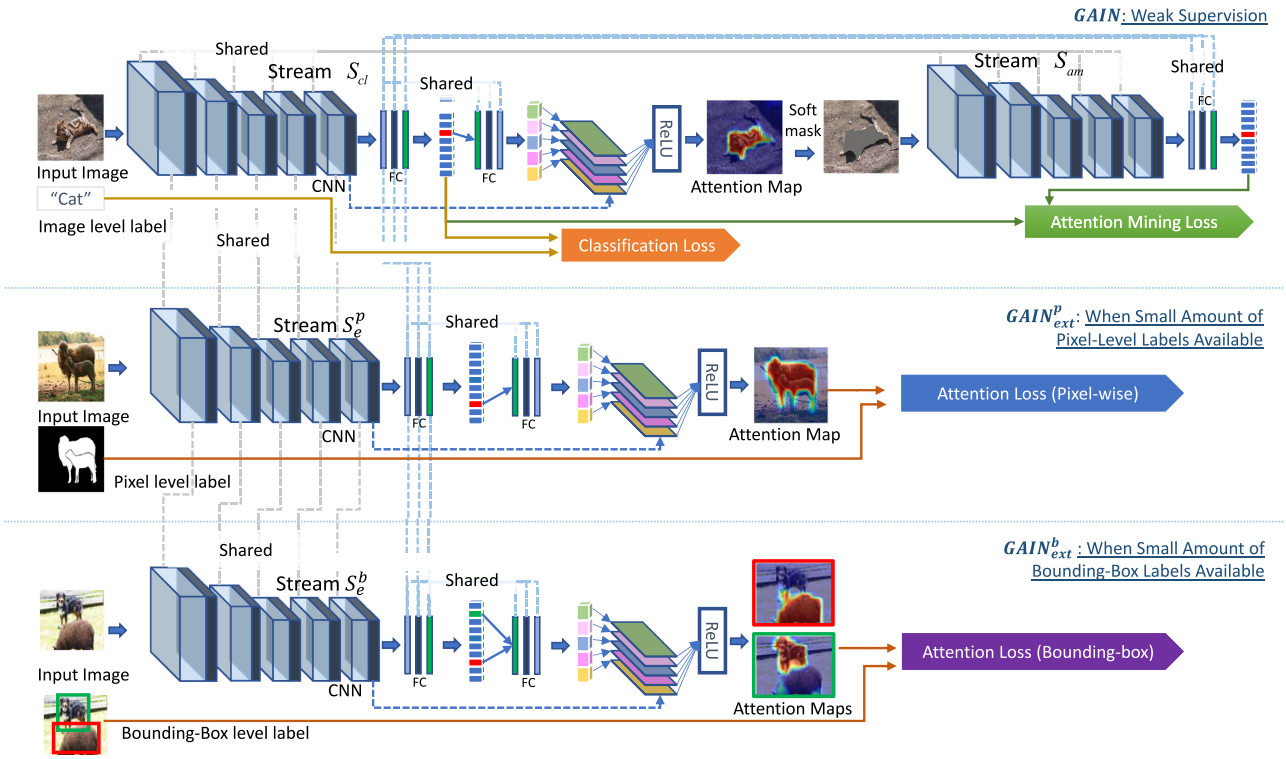


Fig. 3. Framework of the $GAIN_{ext}$. Pixel-level ($GAIN^p_{ext}$) and bounding-box ($GAIN^b_{ext}$) annotations are seamlessly integrated into the GAIN framework to provide direct supervision on attention maps optimizing towards the task of semantic segmentation.

network to learn features robust to dataset bias and improve its generalizability.

3.3 $GAIN^b_{ext}$: Integrating Bounding Box Supervision

Compared with pixel-level segmentation masks, object bounding box annotations are far less expensive, taking only around 1/15 time [29]. Cheaper and easier to define, box annotations could also be integrated as extra supervision to guide the learning process of the network attention maps. They convey useful information to tell network which part of the box content is background and which is foreground object that they should focus on.

Following Section 3.2, we treat the bounding box as a kind of extra supervision. Considering the difference between the bounding box and the pixel-level segmentation mask, we need to modify Stream S^p_e and replace Attention Loss L_p with a new loss function L_{bbox} to provide guidance directly on attention maps. The input of this new stream S^b_e are images with their corresponding bounding box annotations and foreground priors. L_{bbox} encourages the network to pay less attention on the regions out of bounding boxes for class c when recognizing this class, meanwhile, pay more attention on the foreground objects within these bounding boxes. Saliency map can be used here to represent foreground priors. To achieve this, we define L_{bbox} in Eq. (9).

$$L_{bbox} = \frac{1}{n}\sum_{c}\{[(O - B^c) \odot A^c - Z]^2 + [B^c \odot (A^c - S)]^2\}, \tag{9}$$

where B^c is a matrix generated based on the bounding box annotation for class c. Elements of B^c equal to 1 if they are within any bounding box belongs to class c and equal to 0 otherwise. B^c is then resized to be of the same size as that of the attention map A_c. O is a matrix with all elements equal to 1 and Z is a matrix with all elements equal to 0. S is the resized saliency map for current image. O, Z and S all have the same size as that of the attention map A^c. $\odot$ denotes element-wise multiplication.

Again, we are interested in finding out the benefits of using only a very small amount of data with bounding box supervision, which fits perfectly with the $GAIN^b_{ext}$ framework. $GAIN^b_{ext}$ is obtained by replacing stream S^p_e with stream S^b_e. The three streams still share all parameters. A very small amount of bounding box annotations through stream S^b_e can already help to improve the performance with $GAIN^b_{ext}$ (in our experiments with $GAIN^b_{ext}$, only $1\sim 10$ percent of the total labels used in training are bounding box labels).

The final loss function $L_{ext-bbox}$ of $GAIN^b_{ext}$ is defined as follows:

$$L_{ext-bbox} = L_{cl} + \alpha L_{am} + \omega L_{bbox}, \tag{10}$$

where L_{cl} and L_{am} are defined in Section 3.1, and ω is the weighting parameter depending on how much emphasis we want to place on the bounding box supervision (we use $\omega = 10$ in our experiments which is the same as Eq. (8)).

4 Semantic Segmentation Experiments

To verify the efficacy of GAIN, following Sections 3.1, 3.2 and 3.3, we use the weakly supervised semantic segmentation task as the example application. The goal of this task is to classify each pixel into different categories. In the weakly supervised setting, most of recent methods [22], [23], [47] mainly rely on localization cues generated by models trained with only image-level labels and consider other constraints such as object boundaries to train a segmentation network. Therefore,

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

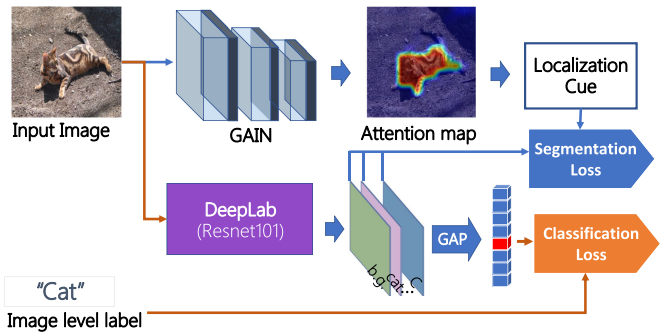


Fig. 4. Structure of the weakly-supervised semantic segmentation framework using fully convolutional network (Deeplab [8] here) as the backbone network. Image-level labels and localization cues obtained based on attention maps of GAIN provide guidance on the training of this segmentation network, which is achieved by optimizing two loss functions jointly. No ground-truth pixel-level annotations are used during the training process of this semantic segmentation network.

the quality of localization cues is the key of these methods' performance.

Compared with attention maps generated by state-of-the-art methods [31], [39], [59] which only locate the most discriminative areas, GAIN guides the network to focus on entire areas representing the class of interest, which can improve the performance of weakly supervised segmentation. To verify this, we first adopt our attention maps to SEC [23], which is one of the state-of-the-art weakly supervised semantic segmentation methods. SEC defines three key constraints: *seed*, *expand* and *constrain*, where *seed* is a module to provide localization cues C to the main segmentation network N such that the segmentation result of N is supervised to match C. Following SEC [23], our localization cue $\ell_x^c \in \{0, 1\}$ for each class c at location x is obtained by applying a thresholding operation to attention maps generated by GAIN: for each per-class attention map, all pixels with a score larger than 20 percent of the maximum score are selected. We apply [30] several times to get background cues and then train the SEC model to generate segmentation results using the same inference procedure, as well as parameters of CRF [24]. According to cues generated by our different GAIN models with different kinds of supervision, we denote the segmentation results as GAIN-SEC, GAIN$^b_{ext}$-SEC and GAIN$^p_{ext}$-SEC accordingly.

In addition to using the existing weakly-supervised semantic segmentation framework for evaluation, we also use a framework similar to [33], [47], [50] to further investigate the benefit of our improved attention maps. As shown in Fig. 4, the framework is built upon Deeplab backbone segmentation network [8] and supervised by two loss functions jointly. One the one hand, due to the improved quality of attention maps of GAIN, they could act as pseudo ground truth to provide guidance on the training of this weakly-supervised semantic segmentation framework. In particular, we obtain localization cue $\ell_x^c \in \{0, 1\}$ for each class c at location x based on attention maps of GAIN following the same way as SEC [23]. Then we define the segmentation loss L_{seg} as follows:

$$L_{seg} = \sum_c J(s_x^c(I), \ell_x^c), \tag{11}$$

where $s_x^c(I)$ is the prediction of the segmentation network of class c at localization x for a input image I. ℓ_x^c is the corresponding localization cue obtained by GAIN. $J(\cdot)$ is the pixel-wise cross entropy loss.

On the other hand, we add a Global Average Pooling layer (GAP) at the last layer of the segmentation network to get the class prediction score for a input image. Based on the understanding that a segmentation network should also have the ability to classify the image well, we use a multi-label classification loss L_{cl} to constrain the learning of the network. The final loss function L_{w-s} is defined as follows:

$$L_{w-s} = L_{cl} + L_{seg}. \tag{12}$$

The trained network then generates segmentation results for evaluation.

4.1 Dataset and Experimental Settings

Datasets and Evaluation Metrics. We evaluate our results on PASCAL VOC 2012 image segmentation benchmark [13], which has 21 semantic classes, including the background. The whole dataset is split into three sets: training, validation, and testing (denoted as train, val, and test) with 1,464, 1,449, and 1,456 images, respectively. Following the common setting [7], [23], we also use the augmented training set provided by [15]. The resulting training set has 10,582 weakly annotated images which we use to train our models. We compare our approach with other methods on both the val. and test sets. For the evaluation metric, we use the standard one for the PASCAL VOC 2012 segmentation—mean intersection-over-union (mIoU).

Implementation Details. We use VGG [42] pretrained from the ImageNet [10] as the basic network for GAIN to generate attention maps. We use PyTorch [1] to implement our models. We set the batch size to 1 and learning rate to 10^{-5}. We use the stochastic gradient descent (SGD) to train the networks and terminate after 35 epochs. For the concern about max-min optimization problem, we have not observed any issue with convergence in our experiments with various datasets and projects. Our total loss decreases around 90 and 98 percent after 1 and 15 epochs respectively. For the weakly-supervised segmentation framework, following the setting of SEC [23], we use the DeepLab-CRFLargeFOV [7], which is a slightly modified version of the VGG network [42]. Implemented using Caffe [20], DeepLab-CRFLargeFOV [7] defines the input size as 321 $\times$ 321 and produces segmentation masks with size of 41 $\times$ 41. Our training procedure is the same as [23] at this stage. We run the SGD for 8,000 iterations with the batch size of 15. The initial learning rate is 10^{-3} and it decreases by a factor of 10 for every 2,000 iterations. For the details about our own weakly-supervised semantic segmentation network described in Section 4, we still use the input size of 321 $\times$ 321. The parameters of back-bone network DeepLab [8] are initialized by ResNet-101 [16] pre-trained on ImageNet [10]. We set the starting learning rate as 0.0005, multiplying it by 0.1 every 2,000 iterations. We use a mini-batch of 10, momentum of 0.3 and weight decay of 0.0005 to train the network with 20,000 iterations.

4.2 Comparison with State-of-the-Art

We compare our methods with other state-of-the-art weakly supervised semantic segmentation methods with image-level labels. Following [47], we separate them into two categories.

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Comparison of Weakly Supervised Semantic Segmentation Methods on PASCAL VOC 2012 *segmentation val.* Set and *segmentation test* Set

Methods	Training Set	*val.* (mIoU)	*test* (mIoU)
Supervision: Purely Image-level Labels			
CCNN [34]	10K W	35.3	35.6
MIL-sppxl [35]	700K W	35.8	36.6
EM-Adapt [33]	10K W	38.2	39.6
DCSM [40]	10K W	44.1	45.1
BFBP [38]	10K W	46.6	48.0
STC [48]	50K W	49.8	51.2
AF-SS [36]	10K W	52.6	52.7
CBTS-cues [37]	10K W	52.8	53.7
TPL [22]	10K W	53.1	53.8
WebS-i2 [21]	10K W	53.4	55.3
PRM [60]	10K W	53.4	-
MEFF [14]	10K W	-	55.6
AE-PSL [47]	10K W	55.0	55.7
SEC [23] (baseline)	10K W	50.7	51.7
GAIN-SEC (ours)	10K W	**55.3**	**56.8**
GAIN (ours)	10K W	**59.4**	**59.6**
Supervision: Bounding box annotations (# Implicitly use bounding box annotations)			
BoxSup$_R$ [9]	10K W + 10K B	52.3	-
WSSL$_R$ [33]	10K W + 10K B	52.5	54.2
WSSL$_S$ [33]	10K W + 10K B	60.6	62.2
GAIN$^b_{ext}$-SEC# (ours)	10K W + 200 B	**56.8**	**57.6**
GAIN$^b_{ext}$-SEC# (ours)	10K W + 1.4K B	**58.0**	**59.2**
GAIN$^b_{ext}$# (ours)	10K W + 200 B	**61.1**	**61.6**
GAIN$^b_{ext}$# (ours)	10K W + 1.4K B	**62.6**	**62.8**
Supervision: Image-level Labels (@ Implicitly use bounding box and pixel-level annotations)			
GAIN$^{b,p}_{ext}$-SEC@	10K W + 100 B + 100 P	**57.8**	**59.3**
GAIN$^{b,p}_{ext}$-SEC@	10K W + 732 B + 732 P	**59.0**	**60.2**
GAIN$^{b,p}_{ext}$@	10K W + 100 B + 100 P	**61.7**	**62.8**
GAIN$^{b,p}_{ext}$@	10K W + 732 B + 732 P	**63.2**	**63.6**
Supervision: Image-level Labels (* Implicitly use pixel-level annotations)			
MIL-seg* [35]	700K W + 1.4K P	40.6	42.0
TransferNet* [17]	27K W + 17K P	51.2	52.1
AF-MCG* [36]	10K W + 1.4K P	54.3	55.5
GAIN$^p_{ext}$-SEC* (ours)	10K W + 200 P	**58.3**	**59.6**
GAIN$^p_{ext}$-SEC* (ours)	10K W + 1.4K P	**60.5**	**62.1**
GAIN$^p_{ext}$* (ours)	10K W + 200 P	**62.2**	**61.9**
GAIN$^p_{ext}$* (ours)	10K W + 1.4K P	**64.1**	**64.4**

"W" denotes image-level labels, "B" denotes bounding box annotations and "P" denotes pixel-level labels. Implicitly use pixel-level supervision *is a protocol we followed as defined in [47], that pixel-level labels are only used in training priors, and only weak labels are used in the training of segmentation framework, e.g., SEC [23] in our case.* Implicitly use bounding box supervision *is a similar protocol.*

For methods that purely use image-level labels, we compare our GAIN-based SEC (denoted as GAIN-SEC in the Table 1) and GAIN with SEC [23] AE-PSL [47], TPL [22], STC [48], MEFF [14], PRM [60] etc. For another group of methods, implicitly using pixel-level supervision means that though these methods train the segmentation networks only with image-level labels, they use some extra technologies that are trained using pixel-level supervision. Our GAIN$^p_{ext}$-based SEC (denoted as GAIN$^p_{ext}$-SEC in the table) and GAIN$^p_{ext}$ lie in this setting because it uses a very small amount of pixel-level labels to further improve the network's attention maps and doesn't rely on any pixel-level labels when training the SEC segmentation network. Other methods in this setting like AF-MCG [59], TransferNet [17] and MIL-seg [35] are included for comparison. For methods that use both image-level labels and some amount of bounding box annotations, BoxSup$_R$ [9], WSSL$_R$ [33] and WSSL$_S$ [33] are included for comparison. The supervision of our method is a kind of implicitly using bounding box supervision. Different from directly using a large amount (10K) bounding box annotations to train the semantic segmentation network, our GAIN$^b_{ext}$-SEC and GAIN$^b_{ext}$ only use a small amount of bounding box labels to train the localization cues. None of these annotations are used to train the semantic segmentation network in the next step. Table 1 shows results on PASCAL VOC 2012 *segmentation val.* set and *segmentation test.* set.

Among the methods purely using image-level labels, our GAIN-based SEC achieves the performance with 55.3 and 56.8 percent in mIoU on these two sets, outperforming the SEC [23] baseline by 4.6 and 5.1 percent. Furthermore, GAIN outperforms AE-PSL [47] by 0.3 and 1.1 percent, and outperforms TPL [22] by 2.2 and 3.0 percent. These two methods are also proposed to improve attention maps to cover more areas of the class of interest. Compared with them, our GAIN makes the attention map trainable without the need to do iterative erasing or combining attention maps from different networks, as proposed in [22], [47]. In addition to using the existing weakly-supervised semantic segmentation framework SEC to evaluate the quality of our improved attention maps, we also show the performance of our proposed framework using attention maps as input cues. GAIN achieves 59.4 and 59.6 percent in mIoU, which further validates the benefit of our improved attention maps.

Our framework also supports to plug in different levels of extra supervision to provide guidance on attention learning of a network. When using bounding box as extra supervision, our GAIN$^b_{ext}$ based SEC further improves performance upon SEC as well as GAIN based SEC because of better attention maps. Based on attention maps using 200 randomly selected bounding box annotations as well as 10K image level weak labels, GAIN$^b_{ext}$-SEC performs 56.8 and 57.6 percent on the *val.* and *test* sets of VOC 2012. When the number of available bounding box annotations increases to 1.4K, GAIN$^b_{ext}$-SEC can achieve 58.0 and 59.2 percent in mIoU on the two sets. When using our semantic segmentation framework with improved attention maps, GAIN$^b_{ext}$ performs 61.1 and 61.6 percent with 200 bounding box annotations on VOC *val.* and *test* sets, and the performance improves to 62.6 and 62.8 percent when the number of bounding box annotations increases to 1.4K. All these bounding box annotations are only used during the attention learning process, none of them are used to train the semantic segmentation model in the next step. This setting helps to validate that the extra supervision can help to further improve the attention maps and guide the attention learning of the network. Besides, for the segmentation task, compared with other methods that use 10K image-level labels and 10K bounding box annotations, our methods achieve better performance, but rely on less bounding box annotations (only 200 or 1.4K bounding box annotations).

By implicitly using pixel-level supervision, our GAIN$^p_{ext}$-based SEC achieves 58.3 and 59.6 percent in mIoU when we use 200 randomly selected images with pixel-level labels (2 percent data of the whole dataset) as the extra supervision. It already performs 4 and 4.1 percent better

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

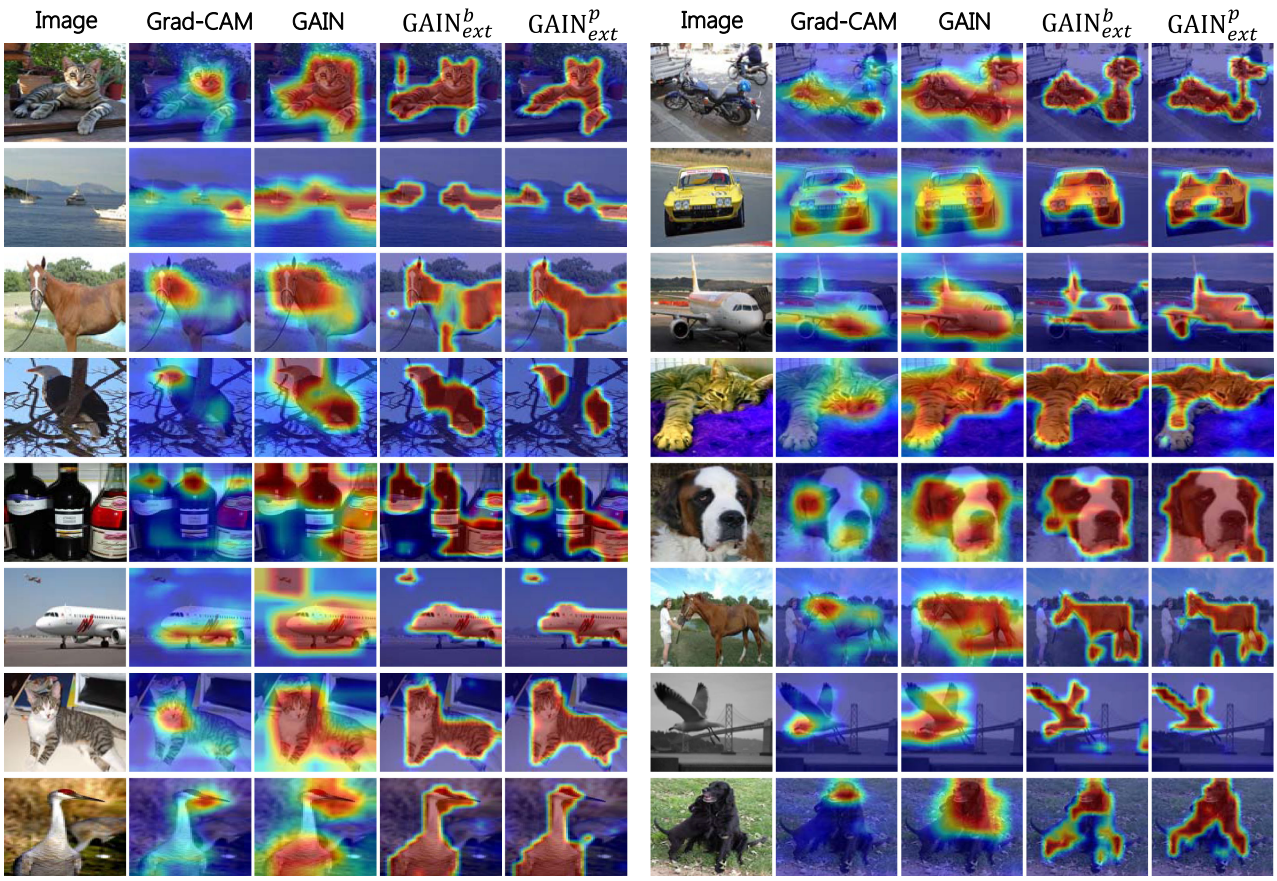


Fig. 5. Qualitative results of attention maps generated by Grad-CAM [39], the proposed GAIN, $GAIN^{b}_{ext}$ and $GAIN^{p}_{ext}$. Here, $GAIN^{b}_{ext}$ uses 200 randomly selected (2 percent) bounding box annotations and $GAIN^{p}_{ext}$ uses 200 randomly selected (2 percent) pixel-level labels as extra supervision to guide the attention learning process. The proposed methods can obtain much better attention maps that cover more complete regions of interest by guided attention learning with different levels of guidance.

than AF-MCG [59], which relies on the MCG generator [3] trained in a fully-supervised way on the PASCAL VOC. After the pixel-level supervision increases to 1,464 images for our $GAIN^{p}_{ext}$-based SEC, the performance jumps to 60.5 and 62.1 percent. When we use the improved attention maps as localization cues to train our own semantic segmentation model, $GAIN^{p}_{ext}$ performs 62.2 and 61.9 percent on the VOC 2012 *val.* and *test* sets using 200 randomly selected pixel-level labels during the guided attention learning. The performance improves to 64.1 and 64.4 percent more extra annotations. Similar to the previous experiment that uses bounding box annotations, none of the pixel-level annotations are used to train the semantic segmentation model in order to validate the improvements are from better attention maps.

We show qualitative results of attention maps generated by GAIN-base methods in Fig. 5, where GAIN covers more areas belonging to the class of interest compared with the Grad-CAM [39]. With only 2 percent of the pixel-level labels, the $GAIN^{p}_{ext}$ covers more complete and accurate areas of the class of interest as well as less background areas around the class of interest (for example, the sea around the ships and the road under the car in the second row of Fig. 5).

Fig. 6 shows some qualitative results of semantic segmentation, indicating that GAIN-based methods help to discover more complete and accurate areas of classes of interest.

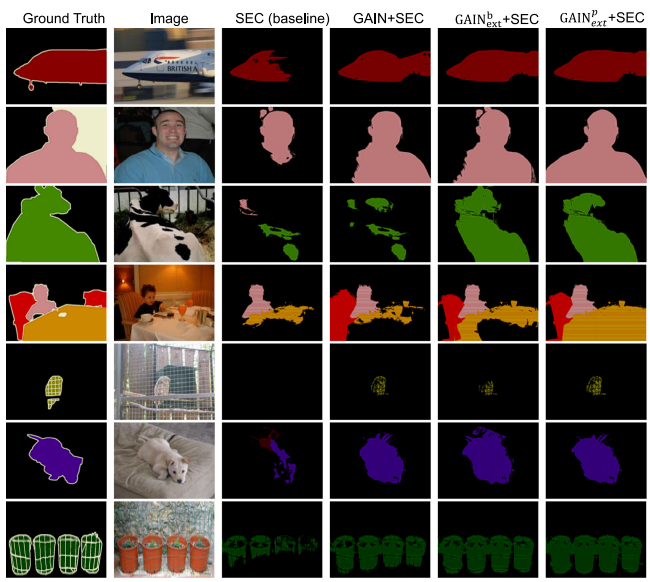


Fig. 6. Qualitative results on PASCAL VOC 2012 *segmentation val.* set. They are generated by SEC (our baseline framework), GAIN-based SEC, $GAIN^{b}_{ext}$-based SEC and $GAIN^{p}_{ext}$-based SEC. Here, $GAIN^{b}_{ext}$-based SEC is based on attention maps generated by $GAIN^{b}_{ext}$ that uses 200 randomly selected (2 percent) bounding box annotations during training. $GAIN^{p}_{ext}$-based SEC is based on attention maps of $GAIN^{p}_{ext}$ that uses 200 randomly selected pixel-level labels as extra supervision to guide the attention learning process. No extra supervision is used during the training of semantic segmentation network, SEC here.

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Comparison of Weakly Supervised Semantic Segmentation Methods on Pascal VOC 2012 *segmentation val.* Set

Methods	b.g.	plane	bike	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	hors.	moto.	pers.	plant	sheep	sofa	train	tv	mIoU
Supervision: Purely Image-level Labels																						
CCNN [34]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	40.7	30.4	36.3	22.2	38.8	36.9	35.3
MIL-sppxl [35]	77.2	37.3	18.4	25.4	28.2	31.9	41.6	48.1	50.7	12.7	45.7	14.6	50.9	44.1	39.2	37.9	28.3	44.0	19.6	37.6	35.0	36.6
DCSM [40]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
BFBP [38]	79.2	60.1	20.4	50.7	41.2	46.3	62.6	49.2	62.3	13.3	49.7	38.1	58.4	49.0	57.0	48.2	27.8	55.1	29.6	54.6	26.6	46.6
STC [48]	84.5	68.0	19.5	60.5	42.5	44.8	68.4	64.0	64.8	14.5	52.0	22.8	58.0	55.3	57.8	60.5	40.6	56.7	23.0	57.1	31.2	49.8
AF-SS [36]	-	61.8	26.8	47.7	27.9	50.2	67.6	59.6	77.5	24.8	51.9	30.5	67.3	52.8	62.9	55.7	37.9	61.4	32.0	50.9	54.1	51.6
CBTS-cues [37]	85.8	65.2	29.4	63.8	31.2	37.2	69.6	64.3	76.2	21.4	56.3	29.8	68.2	60.6	66.2	55.8	30.8	66.1	34.9	48.8	47.1	52.8
AE-PSL [47]	83.4	71.1	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	66.5	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
TPL [22]	82.8	62.2	23.1	65.8	21.1	43.1	71.1	66.2	76.1	21.3	59.6	35.1	70.2	58.8	62.3	66.1	35.8	69.9	33.4	45.9	45.6	53.1
SEC [23] (b.l.)	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.5	52.5	32.5	62.6	32.1	45.4	45.3	50.7
GAIN-SEC	86.9	69.3	29.7	64.0	49.1	51.4	65.8	67.8	73.4	22.0	57.4	20.0	68.7	60.4	63.9	68.1	34.2	63.1	30.0	63.6	52.4	**55.3**
GAIN	87.6	76.7	33.9	74.5	58.5	61.7	75.9	72.9	78.6	18.8	70.8	14.1	68.7	69.6	69.5	71.3	41.5	66.5	16.4	70.2	48.7	**59.4**
Supervision: Bounding box annotations (# Implicitly use bounding box annotations)																						
$GAIN^{b}_{ext}$-SEC-1#	86.8	71.5	30.1	64.5	43.0	51.9	72.6	69.7	75.2	23.7	61.3	34.4	69.4	61.5	65.4	68.2	36.1	65.5	32.5	63.7	46.2	**56.8**
$GAIN^{b}_{ext}$-SEC-2#	87.8	71.8	30.3	66.4	51.3	52.9	74.1	72.1	76.4	23.6	61.3	31.8	71.8	63.1	65.2	69.7	34.3	62.8	32.2	68.0	51.1	**58.0**
$GAIN^{b}_{ext}$-1#	88.5	77.4	35.3	72.7	57.0	62.4	74.5	71.9	77.9	19.4	72.5	17.2	73.9	71.8	71.9	72.9	44.6	72.2	22.8	69.3	56.5	**61.1**
$GAIN^{b}_{ext}$-2#	89.1	78.1	34.4	71.0	64.3	72.3	83.2	75.3	83.4	14.4	73.1	16.2	76.3	70.8	69.4	72.5	43.8	76.0	23.2	70.9	57.2	**62.6**
Supervision: Image-level Labels (@ Implicitly use bounding box and pixel-level annotations)																						
$GAIN^{b,p}_{ext}$-SEC-1@	86.8	71.5	30.1	64.5	43.0	51.9	72.6	69.7	75.2	23.7	61.3	34.4	69.4	61.5	65.4	68.2	36.1	65.5	32.5	63.7	46.2	**57.8**
$GAIN^{b,p}_{ext}$-SEC-2@	88.1	72.5	29.1	66.6	52.4	54.2	76.1	71.6	76.3	23.7	63.3	39.0	72.1	64.3	65.6	70.3	34.7	64.7	34.4	68.2	52.2	**59.0**
$GAIN^{b,p}_{ext}$-1@	88.2	76.8	35.7	73.2	58.5	63.1	72.9	73.2	78.2	20.5	71.8	19.4	74.3	72.5	70.4	73.2	45.1	73.4	23.2	68.9	56.8	**61.7**
$GAIN^{b,p}_{ext}$-2@	89.3	77.8	36.2	73.1	65.6	71.5	82.7	76.4	85.6	16.3	75.4	17.4	74.8	72.3	71.6	70.2	45.3	78.1	24.0	72.3	57.9	**63.2**
Supervision: Image-level Labels (* Implicitly use pixel-level annotations)																						
MIL-seg* [35]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
TransferNet* [17]	85.3	68.5	26.4	69.8	36.7	49.1	68.4	55.8	77.3	6.2	75.2	14.3	69.8	71.5	61.1	31.9	25.5	74.6	33.8	49.6	43.7	52.1
AF-MCG* [36]	-	55.3	27.0	62.0	30.6	56.7	72.8	64.9	79.5	26.7	59.3	31.4	73.0	57.7	63.9	67.4	36.1	68.0	34.6	51.7	38.1	54.3
$GAIN^{p}_{ext}$-SEC-1*	87.8	72.8	30.4	66.9	44.3	49.7	71.5	69.5	77.1	22.7	63.3	45.3	71.3	64.9	65.7	69.8	34.5	65.7	33.0	65.9	51.3	**58.3**
$GAIN^{p}_{ext}$-SEC-2*	88.9	73.7	32.2	69.1	54.2	52.7	75.5	74.0	79.8	24.3	64.7	46.1	73.5	64.8	66.5	72.3	35.3	67.5	33.5	68.3	54.5	**60.5**
$GAIN^{p}_{ext}$-1*	88.9	78.6	36.5	76.0	60.2	67.0	76.8	74.3	81.1	25.0	72.5	16.2	75.3	72.7	71.4	74.4	40.0	72.7	20.5	70.1	55.9	**62.2**
$GAIN^{p}_{ext}$-2*	89.7	82.6	36.0	75.9	63.9	65.9	80.9	74.9	83.0	23.5	76.1	17.9	77.5	75.4	72.6	76.0	40.1	75.7	25.9	73.4	58.9	**64.1**

$GAIN^{b}_{ext}$-SEC-1 and $GAIN^{b}_{ext}$-SEC-2 represent $GAIN^{b}_{ext}$ based SEC that implicitly using 200 and 1464 bounding box annotations respectively. $GAIN^{p}_{ext}$-SEC-1 and $GAIN^{p}_{ext}$-SEC-2 represent $GAIN^{p}_{ext}$ based SEC implicitly using 200 and 1464 pixel-level labels respectively. $GAIN^{b,p}_{ext}$-SEC-1 and $GAIN^{b,p}_{ext}$-SEC-2 represent implicitly using 200 and 1464 combinations of bounding box and pixel-level annotations with a half-and-half ratio respectively. "-1" and "-2" after $GAIN^{b}_{ext}$, $GAIN^{p}_{ext}$, $GAIN^{b,p}_{ext}$ also represent for amount of annotation, which is the same as corresponding SEC-based version.

More Discussion of the $GAIN^{p}_{ext}$. We are interested in finding out the influence of different amount of bounding box annotations or pixel-level labels on the performance. Following the same setting in Section 4.1, we add more randomly selected bounding box annotations or pixel-level labels to further improve attention maps and adopt them in the SEC [23] as well as our weakly-supervised semantic segmentation framework. From the results in Tables 3 and 4, we find that the performance of the semantic segmentation model improves when more extra annotations are provided to train the network that generates attention maps. Again, there are no pixel-level labels used to train the weakly-supervised semantic segmentation framework.

Detailed quantitative results for weakly supervised semantic segmentation experiments including IoU scores for each class of Pascal VOC 2012 *segmentation val., test* set are shown in Tables 2 and 5. We also evaluate performance on VOC 2012 *seg. val.* and *seg. test* datasets without CRF as shown in Table 6.

TABLE 3
Results on PASCAL VOC 2012 *segmentation val.* Set with the Proposed $GAIN^{b}_{ext}$-based SEC and $GAIN^{b}_{ext}$ Implicitly Using Different Amount of Bounding Box Supervision for the Attention Map Learning Process

Method	Training Set	*val.* (mIoU)
$GAIN^{b}_{ext}$-SEC*	10K weak + 200 pixel	56.8
	10K weak + 400 pixel	57.1
	10K weak + 900 pixel	57.3
	10K weak + 1464 pixel	58.0
$GAIN^{b}_{ext}$	10K weak + 200 pixel	61.1
	10K weak + 400 pixel	61.4
	10K weak + 900 pixel	62.2
	10K weak + 1464 pixel	62.6

TABLE 4
Results on PASCAL VOC 2012 *segmentation val.* Set with our $GAIN^{p}_{ext}$-based SEC and $GAIN^{p}_{ext}$ Implicitly Using Different Amount of Pixel-Level Supervision for the Attention Map Learning Process

Method	Training Set	*val.* (mIoU)
$GAIN^{p}_{ext}$-SEC*	10K weak + 200 pixel	58.3
	10K weak + 400 pixel	59.4
	10K weak + 900 pixel	60.2
	10K weak + 1464 pixel	60.5
$GAIN^{p}_{ext}$	10K weak + 200 pixel	62.2
	10K weak + 400 pixel	62.6
	10K weak + 900 pixel	63.3
	10K weak + 1464 pixel	64.1

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

TABLE 5
Comparison of Weakly Supervised Semantic Segmentation Methods on Pascal VOC 2012 *segmentation test* Set

Methods	b.g.	plane	bike	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	hors.	moto.	pers.	plant	sheep	sofa	train	tv	mIoU
Supervision: Purely Image-level Labels																						
CCNN [34]	70.1	24.2	19.9	26.3	18.6	38.1	51.7	42.9	48.2	15.6	37.2	18.3	43.0	38.2	52.2	40.0	33.8	36.0	21.6	33.4	38.3	35.6
MIL-sppxl [35]	74.7	38.8	19.8	27.5	21.7	32.8	40.0	50.1	47.1	7.2	44.8	15.8	49.4	47.3	36.6	36.4	24.3	44.5	21.0	31.5	41.3	35.8
EM-Adapt [33]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
DCSM [40]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
BFBP [38]	80.3	57.5	24.1	66.9	31.7	43.0	67.5	48.6	56.7	12.6	50.9	42.6	59.4	52.9	65.0	44.8	41.3	51.1	33.7	44.4	33.2	48.0
STC [48]	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2
AF-SS [36]	-	58.0	28.0	47.4	25.0	57.5	67.8	55.8	77.1	20.9	54.8	35.5	68.2	57.3	73.1	58.6	40.2	58.2	39.5	44.1	55.1	52.7
CBTS-cues [37]	85.7	58.8	30.5	67.6	24.7	44.7	74.8	61.8	73.7	22.9	57.4	27.5	71.3	64.8	72.4	57.3	37.0	60.4	42.8	42.2	50.6	53.7
TPL [22]	83.4	62.2	26.4	71.8	18.2	49.5	66.5	63.8	73.4	19.0	56.6	35.7	69.3	61.3	71.7	69.2	39.1	66.3	44.8	35.9	45.5	53.8
AE-PSL [47]	85.3	66.9	32.2	77.8	39.1	59.2	63.5	61.4	73.1	17.3	60.9	36.4	70.2	56.8	75.9	52.8	38.7	68.5	34.6	51.2	48.5	55.7
SEC [23] (b.l.)	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
GAIN-SEC	88.0	67.0	30.0	66.3	41.4	60.4	66.8	65.1	71.7	25.5	58.7	22.4	72.3	65.8	68.0	72.0	39.9	64.1	33.4	62.2	52.7	**56.8**
GAIN	88.2	79.3	33.7	67.9	50.5	62.5	76.0	72.2	77.6	20.3	65.8	19.5	72.6	73.0	75.2	71.4	42.4	72.8	21.4	61.5	48.6	**59.6**
Supervision: Image-level Labels (# Implicitly use bounding box annotations)																						
GAIN$^{b}_{ext}$-SEC-1#	87.5	66.3	31.5	63.6	35.7	55.8	73.7	64.8	72.7	27.5	61.5	36.0	72.1	66.0	71.5	70.9	43.5	64.6	40.0	58.6	47.3	**57.6**
GAIN$^{b}_{ext}$-SEC-2#	88.3	67.5	31.8	67.0	44.1	58.6	74.0	67.0	74.6	27.0	62.2	34.5	72.9	67.7	71.7	70.8	42.9	67.1	38.6	63.5	51.8	**59.2**
GAIN$^{b}_{ext}$-1#	89.0	80.4	34.3	69.5	48.8	60.4	75.2	73.6	76.5	20.2	75.0	27.0	77.4	73.8	75.5	74.4	47.6	74.0	28.3	63.0	49.5	**61.6**
GAIN$^{b}_{ext}$-2#	89.4	82.2	33.2	74.5	51.2	70.0	81.7	76.6	81.4	16.0	72.9	20.7	75.3	75.3	76.6	73.1	46.2	75.1	25.9	70.5	51.7	**62.8**
Supervision: Image-level Labels (@ Implicitly use bounding box and pixel-level annotations)																						
GAIN$^{b,p}_{ext}$-SEC-1@	88.1	67.4	31.4	66.1	36.5	57.8	72.6	66.2	75.7	26.9	63.2	43.1	73.4	66.8	70.6	71.1	44.7	67.1	39.7	62.2	54.0	**59.3**
GAIN$^{b,p}_{ext}$-SEC-2@	88.5	71.5	30.3	67.7	42.2	58.3	74.2	66.3	74.8	28.2	64.8	39.5	73.2	70.0	71.8	71.3	45.1	69.0	41.4	63.6	52.7	**60.2**
GAIN$^{b,p}_{ext}$-1@	89.5	82.9	36.2	74.7	50.6	65.4	75.9	77.1	79.2	18.1	70.9	28.5	77.6	73.3	75.6	73.9	45.5	76.2	26.6	65.8	54.4	**62.8**
GAIN$^{b,p}_{ext}$-2@	90.0	83.4	35.8	72.7	52.6	65.3	81.1	76.2	80.0	21.5	70.7	27.1	78.0	76.6	78.9	74.4	43.6	77.3	26.0	72.6	51.9	**63.6**
Supervision: Image-level Labels (* Implicitly use pixel-level annotations)																						
MIL-seg* [35]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
TransferNet* [17]	85.7	70.1	27.8	73.7	37.3	44.8	71.4	53.8	73.0	6.7	62.9	12.4	68.4	73.7	65.9	27.9	23.5	72.3	38.9	45.9	39.2	51.2
AF-MCG* [36]	-	57.9	29.4	60.4	29.0	58.1	70.2	62.3	79.6	23.5	57.1	37.5	75.1	60.7	76.4	67.8	40.2	71.0	40.3	44.0	39.6	55.5
GAIN$^{p}_{ext}$-SEC-1*	88.4	69.2	31.4	64.3	37.2	57.2	72.2	66.4	74.6	26.7	63.0	45.1	74.4	68.1	71.3	71.9	46.0	67.1	41.2	61.9	53.0	**59.6**
GAIN$^{p}_{ext}$-SEC-2*	89.7	72.0	33.1	72.3	45.6	59.9	74.4	67.7	79.1	27.5	63.3	46.9	74.5	69.6	72.7	74.1	45.7	70.4	43.3	67.0	55.6	**62.1**
GAIN$^{p}_{ext}$-1*	89.2	78.9	35.2	73.1	52.1	64.1	76.3	74.7	78.5	22.5	72.3	19.8	77.5	76.5	75.1	74.4	44.2	73.9	27.5	63.9	50.5	**61.9**
GAIN$^{p}_{ext}$-2*	89.9	84.5	35.2	69.6	54.3	62.7	81.8	76.9	80.5	24.9	73.5	32.8	77.5	75.4	77.1	75.4	48.0	76.5	33.8	72.9	49.5	**64.4**

GAIN$^{b}_{ext}$-SEC-1 and GAIN$^{b}_{ext}$-SEC-2 represent GAIN$^{b}_{ext}$ based SEC that implicitly using 200 and 1464 bounding box annotations respectively. GAIN$^{p}_{ext}$-SEC-1 and GAIN$^{p}_{ext}$-SEC-2 represent GAIN$^{p}_{ext}$ based SEC implicitly using 200 and 1464 pixel-level labels respectively. GAIN$^{b,p}_{ext}$-SEC-1 and GAIN$^{b,p}_{ext}$-SEC-2 represent implicitly using 200 and 1464 combinations of bounding box and pixel-level annotations with a half-and-half ratio respectively. "-1" and "-2" after GAIN$^{b}_{ext}$, GAIN$^{p}_{ext}$, GAIN$^{b,p}_{ext}$ also represent for amount of annotation, which is the same as corresponding SEC-based version.

5 Guided Learning with Biased Data

In this section, we design two experiments to verify that our methods have potentials to make the classification network robust to dataset bias and improve its generalization ability by providing guidance on its attention.

5.1 Experiment on Boat Recognition

The attention map can be used to identify bias in datasets [46], we take an example from our observations to explain it. We find that the classification network (like Grad-CAM [39]) trained on Pascal VOC dataset always focuses on the sea and water regions instead of boats when predicting there are boats in an image. It means that there exists bias in the Pascal VOC dataset that ships always appear together with the water in most cases. Therefore, the model learns to detect water rather than the pattern or characteristics to recognize the boats themselves, which limits the generalization ability of the learned model. Though we can build a more balanced dataset based on the observations, our GAIN and GAIN$^{p}_{ext}$ provide another way to make the model learn to be robust to the bias without the need to rebuild the dataset.

TABLE 6
Semantic Segmentation Results without CRF on PASCAL VOC 2012 *segmentation val.* and *test* Sets

Methods	Training Set	*val.*	*test*
SEC [23] w/o. CRF	10K weak	44.8	45.4
GAIN-SEC w/o. CRF	10K weak	50.8	51.8
GAIN$^{b}_{ext}$-SEC w/o. CRF	10K weak + 1464 bbox	52.1	52.8
GAIN$^{p}_{ext}$-SEC w/o. CRF	10K weak + 1464 pixel	54.8	55.7

Numbers shown are mIoU.

Experimental Setting Details. To verify this, we construct a test dataset, namely "*Biased Boat*" dataset, containing two categories of images: boat images without sea or water; and sea or water images without boats. We collected 50 images from Internet for each scenario, resulting in 100 images in total. For models, we use the VGG [42] pretrained from the ImageNet [10] as the basic network for Grad-CAM (basic classification network), our GAIN and GAIN$^{p}_{ext}$. We use the stochastic gradient descent to train the networks and terminate after 35 epochs on the training set of Pascal VOC provided by [15] which includes 10,582 images. For our GAIN$^{p}_{ext}$, a small amount of data in the training set have both image-level and pixel-level labels. Following the settings in Section 4.1, we provide 200, 400 and 1,464 randomly chosen pixel-level labels to train our GAIN$^{p}_{ext}$ separately. In these randomly chosen images, there are 9, 23 and 78 images including the boat class,

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

TABLE 7
Results Comparison of VGG with Our GAIN and GAIN$^p_{ext}$ Tested on the *Biased Boat* Dataset for Classification Accuracy

Test set	VGG	GAIN	GAIN$^p_{ext}$ (# of PL)		
			9	23	78
VOC val.	83%	90%	93%	93%	94%
Boat without water	42%	48%	64%	74%	84%
Water without boat	30%	62%	68%	76%	84%
Overall	36%	55%	66%	75%	84%

PL labels *denotes pixel-level labels of* boat *used in the training which are randomly chosen.*

which is used to represent different GAIN$^p_{ext}$ models. The quantitative results are reported in Table 7 including the accuracy on the whole bias dataset as well as for each scenario. We also show qualitative results in Fig. 7. Attention maps are shown when there is boat being recognized.

Discussion and Analysis. From the results it can be seen that with VGG training on VOC 2012, the network is having trouble predicting whether a boat is in the image in both of the two scenarios with 36 percent overall accuracy. In particular it generates positive prediction incorrectly on images with only water 70 percent of the time, indicating that "water" is considered as one of the most prominent feature characterizing "boat" by the network. Using GAIN with only image-level supervision, the overall accuracy on our *Biased Boat* dataset has been improved to 55 percent, with significant improvement (32 percent higher in accuracy, almost half the error rate) on the scenario of "water without boat". This could be attributed to that GAIN is able to teach the learner to capture all relevant parts of the target object, in this case, both the boat itself and the water surrounding it in the image. Hence when there is no boat but water in the image, the network is more likely to generate a negative prediction. However with the help of self-guidance, GAIN still is unable to decouple boat from water due to the biased training data, i.e., the learner is unable to move its attention away from water. That is the reason why only 6 percent improvement on accuracy is observed in the scenario of "boat without water". On the other hand with GAIN$^p_{ext}$ training a with small amount of pixel-level labels, similar levels of improvements are observed in both of the two scenarios. From Table 7 it can be seen that with only 9 pixel-level labels for "boat", GAIN$^p_{ext}$ obtained an overall accuracy of 66 percent on our *Biased Boat* dataset, a 11 percent improvement compared to GAIN with only self-guidance. In particular significant improvement is observed in the scenario of boats without water (16 percent increase on accuracy compared to GAIN, about 30 percent reduction of error). With 78 pixel-level labels for "boat" used in training, GAIN$^p_{ext}$ is able to obtain 84 percent of accuracy on our *Biased Boat* dataset and performance on both of the two scenarios converged. The reasons behind these results could be that pixel-level labels are able to precisely tell the learner what are the relevant features, components or parts of the target objects hence the actual boats in the image can be decoupled from the water. This again supports that by directly providing extra guidance on attention maps, the negative impact from the bias in training data can be greatly alleviated.

5.2 Experiment about Recognizing the Orientation of an Industrial Camera

The second experiment is designed for a challenging case to further verify the model's generalization ability. This industrial application aims to recognize the orientation of the camera. We define two orientation categories for the industrial camera which is highly symmetric in shape. As shown in Fig. 8, only the texture such as the location of the gap and small markers on the surface of the camera are critical to distinguish their orientations.

Fig. 7. Qualitative results generated by Grad-CAM [39], our GAIN and GAIN$^p_{ext}$ on the *Biased Boat* dataset. **-#** denotes the number of pixel-level labels of *boat* used in the training which were randomly chosen from VOC 2012. Attention map corresponding to *boat* shown only when there are boats recognized.

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

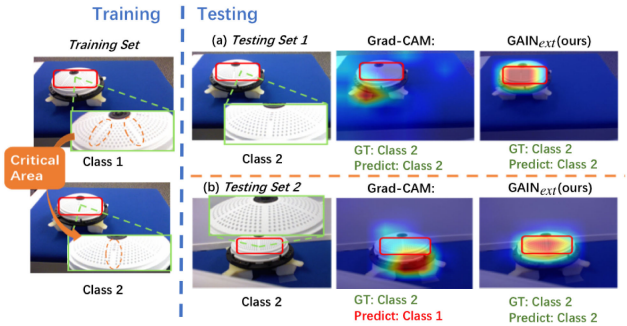


Fig. 8. Datasets and qualitative results of our toy experiments. The critical areas are marked with red bounding boxes in each image. *GT* means ground truth orientation class label.

Experimental Setting Details. We collect two datasets that have different viewpoints and backgrounds. One is divided into two sets: *Training Set* and *Testing Set 1*. There are 350 images for each orientation category in the *Training Set* resulting in 700 images in total for training. *Testing Set 1* includes 100 images sharing same distribution with the training set. For *Testing Set 2*, there are still 50 images per orientation, but the background and viewpoint are different from *Testing Set 1* and *Training Set*. We train VGG-based classification networks [42] without attention guidance (like Grad-CAM [39]) and our GAIN$^p_{ext}$ on the *Training Set*. GAIN$^p_{ext}$ has two streams classification stream S_{cl} and external stream S^b_e. The input images of stream S^b_e include both the image-level labels and bbox labels. Here manually drawn bounding boxes (20 for each classes taking up only 5 percent of the whole training data) on the critical areas are used as external supervision. These bounding boxes are then converted to pixel-level masks to guide the network focus on the critical areas.

Discussion and Analysis. At testing stage, though the Grad-CAM can correctly classify (close to 100 percent accuracy) the images in the *Testing Set 1* where the camera viewpoint and background are very similar to the *Training Set*, it only gets random guess results (close to 50 percent accuracy) on *Testing Set 2* where images are taken from a different viewpoint with different background. This is due to the fact that there is severe bias in *Training Set* and the learner fails to capture the right features (critical area as noted in Fig. 8) to separate the two classes. On the contrary, using GAIN$^p_{ext}$ with a small amount of images with bounding-box labels, the network is able to focus its attention on the area specified by the bounding box labels hence better generalization is observed when testing with *Testing Set 2*. Although the camera viewpoint and scene background are quite different, the learner can still correctly identify the critical area on the camera in the image as shown in the last column second row in 8. Hence it correctly classifies all images in both *Testing Set 1* and *Testing Set 2*.

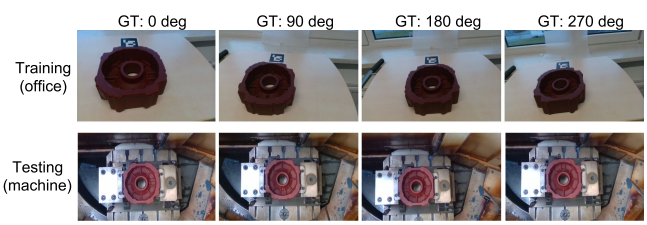


Fig. 9. Example images of the data used in the coarse orientation classification of an industrial workpiece.

TABLE 8
Performance Comparison between the DenseNet [18] and DenseNet-based GAIN$^p_{ext}$

Method	training set	classification accuracy (%)
DenseNet [18]	6,118 weak	42.30
GAIN$^p_{ext}$	6,118 weak + 60 bbox	58.96
GAIN$^p_{ext}$	6,118 weak + 120 bbox	73.00

All the methods are trained with office data but evaluated on 2059 in-machine testing data.

5.3 Experiment about Coarse Orientation Detection of an Industrial Workpiece

We use this experiment as an example to demonstrate that GAIN$^b_{ext}$ can improve the generalization ability of the trained model. In this experiment, our task is to classify the input image of a particular industrial workpiece into one of the four coarse orientations (i.e., 0, 90, 180, and 270 degrees). We are given 6,118 training images acquired in an office environment. However, the trained model needs to be deployed in the factory and tested on 2,059 in-machine testing images. Fig. 9 displays some example images in the training and testing datasets, showing clear difference between these two datasets in terms of scale, viewpoint, background, etc.

In our implementation, we change the base network architecture of the GAIN$^b_{ext}$ from VGG to the DenseNet [18] and compare the classification accuracy of the adapted GAIN$^b_{ext}$ with that of a DenseNet [18] trained with image-level labels. The result is summarized in Table 8, where GAIN$^b_{ext}$ shows better accuracy and generalization ability compared with the DenseNet by using about 1∼2 percent of extra bounding box annotations. Table 8 also shows that using more bounding box annotations with GAIN$^b_{ext}$ achieves better performance, which is consistent with Table 1. Fig. 10 displays some example attention maps corresponding to the three methods in Table 8, showing that using extra and more bounding box annotations can guide the attention of the network to focus more on the workpiece, regardless of the correctness of prediction. Tables 1 and 8 together support that GAIN$^b_{ext}$ outperforms the comparable methods regardless of the tasks (classification and segmentation) and the base network architecture (VGG and DenseNet).

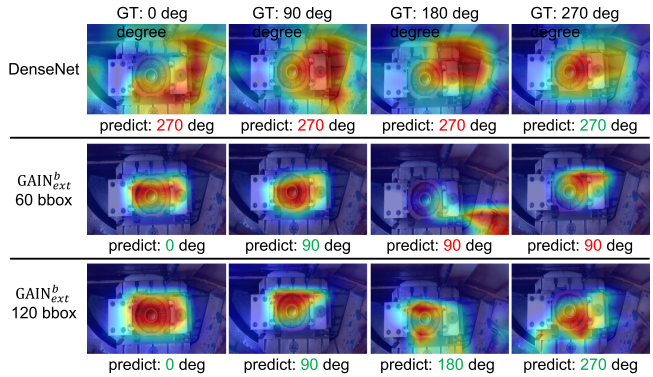


Fig. 10. Some example attention maps corresponding to the three methods in Table 8. These attention maps show that using extra and more bounding box annotations encourages the network to predict based on the workpiece instead of the background.

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

6 Conclusions

In this paper, we propose a framework that can provide guidance directly on the attention maps of a deep convolutional neural network. The network is guided to focus on the regions we expect without changing the network structure or learning extra parameters. This is achieved by making the attention maps not an afterthought, but a first-class citizen during end-to-end training. We validate the effectiveness of the proposed method considering two main roles of these maps. First, when serving as localization priors for tasks like weakly-supervised semantic segmentation, our attention maps can be guided by self supervision or available extra supervision during training, which leads to improvement by covering more complete regions of class of interest. Extensive experiments on the PASCAL VOC 2012 benchmark demonstrate that the proposed method confidently outperforms the state of the art without the need for recursive processing during run time.

Since attention maps are related to the network's response given specific patterns and tasks it was trained for, providing guidance on attention maps can be understand as a regularization for the network learning. As one benefit of this we are able to control the network attention explicitly and can put manual effort in providing minimal supervision of attention rather than re-balancing the dataset when the network suffers from dataset bias. While it may not always be clear how to manually balance datasets to avoid bias, it is usually straightforward to guide attention to the regions of interest. We design several experiments using data from standard benchmark as well as real industrial application to validate this idea. We observe that our explicit guided attention model can help to improve the generalization performance.

In the future it may be illuminating to deploy our method on other tasks related to localization, such as weakly-supervised object detection, object localization, sound-to-visual localization, action localization etc. Besides, constraints based on the principle that learning concepts by focusing on right regions are implemented in this work, which are helpful to improve the generalization ability of the network. Based on our exploration for the relationship between network decision and its attention, we expect more future work can attempt to formulate different constraints on this relationship according to the requirements of specific tasks. Our insight and framework have already inspire several further explorations in other applications. [58] extends our GAIN framework to Siamese network and adopts our Attention Mining Loss L_{am} to the person re-identification module, helping the network generate complete attentive regions in person images. Experiments show a complete attention map can help to improve the performance of person re-identification. [45] validates that a complete attention map can help to improve the model's robustness to adversarial attacks. [11] demonstrates that penalizing the changes in classifier's attention maps helps to retain information of the base classes, as new classes are added, which can help to boost the performance of incremental learning.

Acknowledgments

This paper is based primarily on the work done during Kunpeng Li's internship at Siemens Corporate Technology. This research is supported in part by the NSF IIS Award 1651902 and U.S. Army Research Office Award W911NF-17-1-0367.

References

[1] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," *NIPS-W*, 2017

[2] 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.*, 2018, pp. 4981–4990.

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

[4] S. A. Bargal, A. Zunino, D. Kim, J. Zhang, V. Murino, and S. Sclaroff, "Excitation backprop for RNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1440–1449.

[5] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[6] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu et al., "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2956–2964.

[7] 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 *Proc. Int. Conf. Learn. Representations*, 2015, https://arxiv.org/abs/1412.7062

[8] 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.

[9] 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.*, 2015, pp. 1635–1643.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[11] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5138–5146.

[12] A. Diba, V. Sharma, A. M. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5131–5139.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[14] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1277–1286.

[15] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 991–998.

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

[17] S. Hong, J. Oh, H. Lee, and B. Han, "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3204–3212.

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

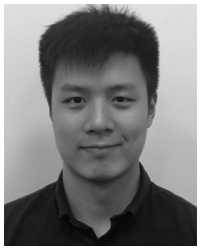
[19] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," in *Proc. Int. Conf. Learn. Representations*, 2018, https://openreview.net/pdf?id=HyzbhfWRW

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[21] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Webly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1705–1714.

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

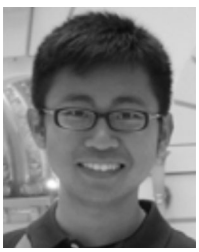
[22] D. Kim, D. Cho, D. Yoo, and I. So Kweon, "Two-phase learning for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3554–3563.
[23] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 695–711.
[24] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[25] K. Li, Z. Ding, K. Li, Y. Zhang, and Y. Fu, "Support neighbor loss for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1492–1500.
[26] K. Li, Y. Kong, and Y. Fu, "Multi-stream deep similarity learning networks for visual tracking," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2166–2172.
[27] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9223.
[28] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Representations*, 2014, https://arxiv.org/abs/1312.4400
[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[30] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 678–686.
[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[32] 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.*, 2015, pp. 685–694.
[33] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a DCNN for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1742–1750.
[34] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1796–1804.
[35] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1713–1721.
[36] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 90–105.
[37] 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.*, 2017, pp. 7282–7291.
[38] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 413–432.
[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[40] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 218–234.
[41] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Representations Workshop*, 2014, https://arxiv.org/abs/1312.6034
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, https://arxiv.org/abs/1409.1556
[43] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3544–3553.
[44] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. Int. Conf. Learn. Representations Workshop*, 2015, https://arxiv.org/abs/1412.6806
[45] A. Subramanya, V. Pillai, and H. Pirsiavash, "Towards hiding adversarial examples from network interpretation," *arXiv preprint arXiv: 1812.02843*, 2018.
[46] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.
[47] 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.*, 2017, pp. 6488–6496.
[48] 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.
[49] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang, "TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 454–470.
[50] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7268–7277.
[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[52] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 543–559.
[53] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1325–1334.
[54] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 610–625.
[55] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 610–625.
[56] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[57] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3549–3557.
[58] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5735–5744.
[59] 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.*, 2016, pp. 2921–2929.
[60] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3791–3800.



Kunpeng Li received the BEng degree in information engineering from the South China University of Technology, China. He is now working toward the PhD degree in Computer Engineering, Northeastern University, Boston, MA. He was a research intern with Adobe Research in San Jose, CA and Siemens Corporate Research in Princeton, NJ. His research interests include scene understanding, vision and language, video analytics, visual recognition and deep learning. He is a student member of the IEEE.

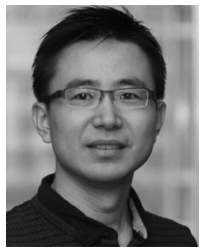
Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.

Ziyan Wu received the BS and MS degrees in measurement technology and instruments, both from Beihang University in China, in 2006 and 2009, respectively, and the PhD degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, in 2014. He is a senior key expert scientist of the Vision Technologies and Solutions Group at Siemens Corporation, Corporate Technology, in Princeton, NJ. He was affiliated with the DHS Center of Excellence on Explosives Detection, Mitigation and Response (ALERT). His research interests include 3D object recognition, scene understanding, video surveillance, deep learning and augmented reality. He organized the VIEW workshop in conjunction with CVPR in 2016, the CVPR Industry EXPO Spotlight in 2017, and the Vision with Biased or Scarce Data workshop in conjunction with CVPR in 2018 and 2019. He is a member of the IEEE.



Kuan-Chuan Peng received the BS degree in electrical engineering, the MS degree in computer science from National Taiwan University, in 2009 and 2012 respectively, and the PhD degree in electrical and computer engineering from Cornell University, in 2016. He is a staff scientist with the Vision Technologies and Solutions Group, Siemens Corporation, Corporate Technology in Princeton, NJ. His research interests include deep learning, abstract tasks, and fundamental computer vision and machine learning problems. He organized the Vision with Biased or Scarce Data workshop in conjunction with CVPR, in 2018 and 2019. He is a member of the IEEE.

Jan Ernst received the PhD degree from the University of Erlangen-Nuremberg in Erlangen, Germany. He is a Principal Key Expert scientist of the Simulation and Digital Twin Technical Field at Siemens Corporation, Corporate Technology in Princeton, NJ. Before becoming the Principal Key Expert scientist, he has been in the position of Senior Key Expert Scientist, Research Group head, and project manager at Siemens. He is a certified R&D Project Management Professional. He has 20 years of experiences in the field of computer vision and machine learning. He is a member of the IEEE.



Yun Fu (S'07-M'08-SM'11-F'19) received the BEng degree in information engineering the MEng degree in pattern recognition and intelligence systems from Xian Jiaotong University, China, respectively, the MS degree in statistics and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University since 2012. His research interests are machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an associate editor of the *IEEE Transactions on Neural Networks and Leaning Systems (TNNLS)*. He is fellow of the IEEE, IAPR, OSA and SPIE, a Lifetime Distinguished Member of ACM, Lifetime Member of AAAI and Institute of Mathematical Statistics, member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS and Beckman Graduate Fellow during 2007-2008.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: UNIVERSITY OF NEVADA RENO. Downloaded on November 15,2023 at 05:27:23 UTC from IEEE Xplore. Restrictions apply.