# Tell Me Where to Look: Guided Attention Inference Network

Kunpeng Li[1], Ziyan Wu[3], Kuan-Chuan Peng[3], Jan Ernst[3] and Yun Fu[1,2]
[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA
[2]College of Computer and Information Science, Northeastern University, Boston, MA
[3]Siemens Corporate Technology, Princeton, NJ
{kunpengli,yunfu}@ece.neu.edu, {ziyan.wu, kuanchuan.peng, jan.ernst}@siemens.com

## Abstract

*Weakly supervised learning with only coarse labels can obtain visual explanations of deep neural network such as attention maps by back-propagating gradients. These attention maps are then available as priors for tasks such as object localization and semantic segmentation. In one common framework we address three shortcomings of previous approaches in modeling such attention maps: We (1) make attention maps an explicit and natural component of the end-to-end training for the first time, (2) provide self-guidance directly on these maps by exploring supervision from the network itself to improve them, and (3) 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. We clearly surpass the state-of-the-art on PASCAL VOC 2012 test and val. sets. 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 weakly supervised learners to improve their generalization performance.*

## 1. Introduction

Weakly supervised learning [3, 26, 33, 36] has recently gained much attention as a popular solution to address labeled data scarcity in computer vision. Using only image level labels for example, one can obtain attention maps for a given input with back-propagation on a Convolutional Neural Network (CNN). These maps relate to the network's response given specific patterns and tasks it was trained for. The value of each pixel on an attention map reveals to what extent the same pixel on the input image contributes to the final output of the network. It has been shown that one can extract localization and segmentation information from such attention maps without extra labeling effort [39].

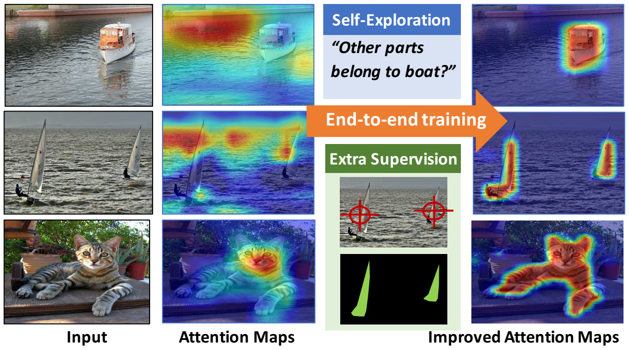

Figure 1. The proposed Guided Attention Inference Network (GAIN) makes the network's attention on-line trainable and can plug in different kinds of supervision directly on attention maps in an end-to-end way. We explore the self-guided supervision from the network itself and propose GAIN$_{ext}$ when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

However, supervised by only classification loss, attention maps often only cover small and most discriminative regions of object of interest [11, 28, 39]. While these attention maps can still serve as reliable priors for tasks like segmentation [12], having attention maps covering the target foreground objects as complete as possible can further boost the performance. To this end, several recent works either rely on combining multiple attention maps from a network via iterative erasing steps [31] or consolidating attention maps from multiple networks [11]. Instead of passively exploiting trained network attention, we envision an end-to-end framework with which task-specific supervision can be directly applied to attention maps during training stage.

On the other hand, as an effective way to explain the network's decision, attention maps can help to find restrictions of the training network. For instance in an object categorization task with only image-level object class labels, we may encounter a pathological bias in the training data when the foreground object incidentally always correlates with the same background object (also pointed out in [24]). Figure 1 shows the example class "boat" where there may be bias towards water as a distractor with high correlation. In this case the training has no incentive to focus attention only on the foreground class and generalization performance may suffer when the testing data does not have the same correlation ("boats out of water"). While there have been attempts to remove this bias by re-balancing the training data, we instead propose to model the attention map *explicitly* as part of the training. 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 re-balancing the data set. While it may not always be clear how to manually balance data sets to avoid bias, it is usually straightforward to guide attention to the regions of interest. We also observe that our explicit self-guided attention model already improves the generalization performance even without extra supervision.

Our contributions are: (a) A method of using supervision directly on attention maps during training time while learning a weakly labeled task; (b) A scheme for self-guidance during training that forces the network to focus attention on the object holistically rather than only the most discriminative parts; (c) Integration of direct supervision and self-guidance to seamlessly scale from using only weak labels to using full supervision in one common framework.

Experiments using semantic segmentation as task of interest show that our approach achieves mIoU 55.3% and 56.8%, respectively on the PASCAL VOC 2012 segmentation *test* and *val* sets. It also confidently surpasses the comparable state-of-the-art when limited pixel-level supervision is used in training with an mIoU of 60.5% and 62.1% respectively.

## 2. Related work

Since deep neural networks have achieved great success in many areas [7, 34, 35, 37], various methods have been proposed to explain this black box [3, 26, 33, 38]. Visual attention is one way that tries to explain which region of the image is responsible for the network's decision. In [26, 29, 33], error back propagation based methods are applied to visualize regions that are helpful for predicting a class. [3] proposes a feedback method to capture the top-down neural attention, which can be used to show task-related regions. CAM [39] shows that the average pooling layer can help to generate attention maps representing task relevant regions than fully-connected layers. Inspired by a top-down visual attention model for human, [36] proposes a new back propagation method, Excitation Backprop, to pass along signals from top to down in the network hierarchy. Recently, Grad-CAM [24] extends CAM [39] to many different available architectures for tasks like image captioning and VQA, which helps to explain model decisions. Different from all these methods that are trying to explain the network, we first time build up an end-to-end model to provide supervision directly on these explanations, specifically network's attention here. We validate that the supervision can guide the network to focus on the regions we expect, which will benefit the corresponding visual task.

Many methods heavily rely on the location information provided by the network's attention. Learning from only the image-level labels, attention maps of a trained classification network can be used for weakly-supervised object localization [17, 39], scene segmentation [12] etc. However, only trained with classification loss, the attention map only covers small and most discriminative regions of the object of interest, which deviates from the requirement of these tasks that needs to localize dense, interior and complete regions. To mitigate this gap, [28] proposes to hide patches in a training image randomly, forcing the network to seek other relevant parts when the most discriminative part is hidden. This approach can be considered as a way to augment the training data, and it has strong assumption on the size of foreground objects (i.e., the object size vs. the size of the patches). [31] uses the attention map of a trained network to erase the most discriminative regions of the original input image. They repeat this erase and discover action to the erased image for several steps and combine attention maps of each step to get a more complete attention map. Similarly, [11] uses a two-phase learning strategy and combines attention maps of the two networks to get a more complete region for the object of interest. In the first step, a conventional fully convolutional network (FCN) [16] is trained to find the most discriminative parts of an image. Then these most salient parts are used to suppress the feature map of the second network to force it to focus on the next most important parts. However, these methods either rely on combinations of attention maps of one trained network for different erased steps or attentions of different networks. The single network's attention still only locates on the most discriminative region. Our proposed GAIN model is fundamentally different from the previous approaches. Since our models can provide supervision directly on network's attention in an end-to-end way, which can not be done by all the other methods [11, 24, 28, 31, 36, 39], we design different kinds of loss functions to guide the network to focus on the whole object of interest. Therefore, we do not need to do erasing for several times or combine attention maps. The attention of our single trained network is already more complete and improved.

Identifying bias in datasets [30] is another important usage of the network attention. [24] analyzes the location of attention maps of a trained model to find out the dataset bias, which helps them to build a better unbiased dataset. However, in practical applications, it is hard to remove all the biases of the dataset and time-consuming to build a new dataset. How to guarantee the generalization ability of the

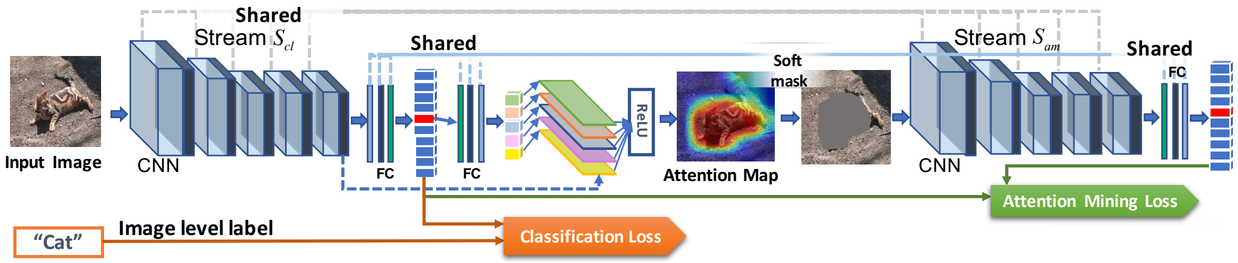

Figure 2. GAIN has two streams of networks, $S_{cl}$ and $S_{am}$, sharing parameters. $S_{cl}$ aims to find out regions that help to recognize the object and $S_{am}$ tries to make sure all these regions contributing to this recognition have been discovered. The attention map is on-line generated and trainable by the two loss functions jointly.

learned network is still challenging. Different from the existing methods, our model can fundamentally solve this problem by providing 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 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 (GAIN), 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 [11, 24, 28, 31, 36, 39]. 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 constrains directly on the attention map in a regularized bootstrapping fashion. As shown in Figure 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 [24], 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 $l$-th 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 average pooling layer [14] to obtain the neuron importance weights $w^c_{l,k}$ as defined in Eq. 1.

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

where $\mathrm{GAP}(\cdot)$ 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:

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

where $l$ is the representation from the last convolutional layer whose features have a good balance between detailed spatial information and high-level semantics [26].

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 $c$.

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

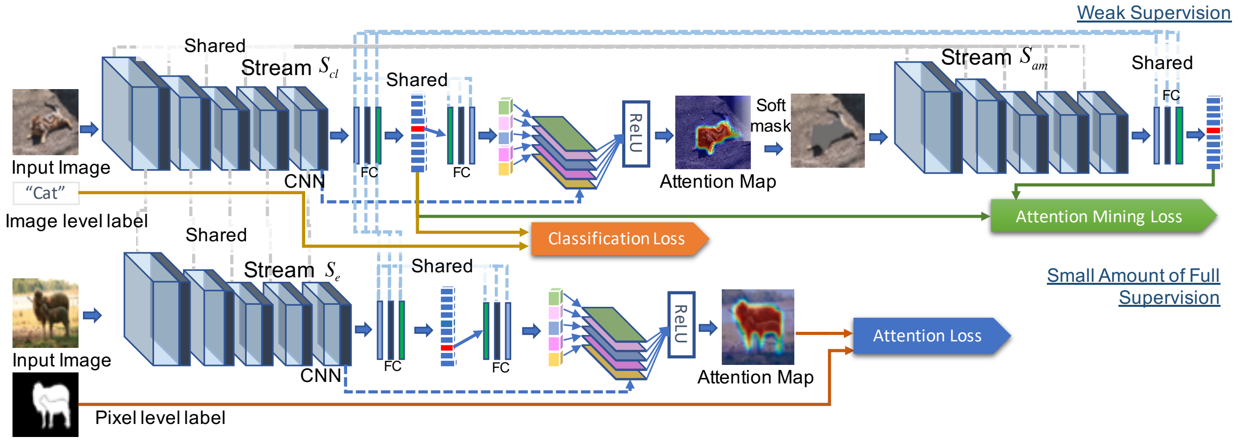

Figure 3. Framework of the GAIN$_{ext}$. Pixel-level annotations are seamlessly integrated into the GAIN framework to provide direct supervision on attention maps optimizing towards the task of semantic segmentation.

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.

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

where $\boldsymbol{\sigma}$ is the threshold matrix whose elements all equal to $\sigma$. $\omega$ is the scale parameter ensuring $T\left(A^c\right)_{i,j}$ approximately equals to 1 when ${A^c}_{i,j}$ is larger than $\sigma$, 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. $\alpha$ 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 Sec. 4.

### 3.2. GAIN$_{ext}$: integrating extra 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$_{ext}$, which can seamlessly integrate extra supervision in our weakly supervised learning framework. We demonstrate using GAIN$_{ext}$ to improve the weakly-supervised semantic segmentation task as shown in Sec. 4. Furthermore, we can also apply GAIN$_{ext}$ to guide the network to learn features robust to dataset bias and improve its generalizability when the testing data and training data are drawn from very different distributions.

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

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

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$_{ext}$ framework shown in Figure 3, where we add an external stream $S_e$, and these three streams share all parameters. Input images of stream $S_e$ include both image-level labels and pixel-level segmentation masks. One can use only very small amount of pixel-level labels through stream $S_e$ to already gain performance improvement with GAIN$_{ext}$ (in our experiments with GAIN$_{ext}$, only 1∼10% 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, $L_{ext}$, of GAIN$_{ext}$ is defined as follows:

$$L_{ext} = L_{cl} + \alpha L_{am} + \omega L_e, \tag{8}$$

where $L_{cl}$ and $L_{am}$ are defined in Sec. 3.1, and $\omega$ 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$_{ext}$ 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 $L_e$ to guide the network to focus on areas critical to the task of interest. In Sec. 5, we show an example of such modification to guide the network to learn features robust to dataset bias and improve its generalizability. In that case, extra supervision is in the form of bounding boxes.

## 4. Semantic segmentation experiments

To verify the efficacy of GAIN, following Sec. 3.1 and 3.2, 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 [11, 12, 31] 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, the quality of localization cues is the key of these methods' performance.

Compared with attention maps generated by state-of-the-art methods [16, 24, 39] 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 adopt our attention maps to SEC [12], which is one of the state-of-the-art weakly supervised semantic segmentation methods. Following SEC [12], our localization cues are 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% of the maximum score are selected. We apply [15] 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 [13].

| Methods | Training Set | *val.* (mIoU) | *test* (mIoU) |
|---|---|---|---|
| Supervision: Purely Image-level Labels | | | |
| CCNN [19] | 10K weak | 35.3 | 35.6 |
| MIL-sppxl [20] | 700K weak | 35.8 | 36.6 |
| EM-Adapt [18] | 10K weak | 38.2 | 39.6 |
| DCSM [25] | 10K weak | 44.1 | 45.1 |
| BFBP [23] | 10K weak | 46.6 | 48.0 |
| STC [32] | 50K weak | 49.8 | 51.2 |
| AF-SS [21] | 10K weak | 52.6 | 52.7 |
| CBTS-cues [22] | 10K weak | 52.8 | 53.7 |
| TPL [11] | 10K weak | 53.1 | 53.8 |
| AE-PSL [31] | 10K weak | 55.0 | 55.7 |
| SEC [12] (baseline) | 10K weak | 50.7 | 51.7 |
| **GAIN (ours)** | 10K weak | **55.3** | **56.8** |
| Supervision: Image-level Labels<br>(* Implicitly use pixel-level supervision) | | | |
| MIL-seg* [20] | 700K weak + 1464 pixel | 40.6 | 42.0 |
| TransferNet* [9] | 27K weak + 17K pixel | 51.2 | 52.1 |
| AF-MCG* [21] | 10K weak + 1464 pixel | 54.3 | 55.5 |
| **GAIN$_{ext}$* (ours)** | 10K weak + 200 pixel | **58.3** | **59.6** |
| **GAIN$_{ext}$* (ours)** | 10K weak + 1464 pixel | **60.5** | **62.1** |

Table 1. Comparison of weakly supervised semantic segmentation methods on PASCAL VOC 2012 *segmentation val.* set and *segmentation test* set. **weak** denotes image-level labels and **pixel** denotes pixel-level labels. *Implicitly use pixel-level supervision* is a protocol we followed as defined in [31], 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 [12] in our case.

### 4.1. Dataset and experimental settings

**Datasets and evaluation metrics.** We evaluate our results on PASCAL VOC 2012 image segmentation benchmark [6], which includes 20 foreground classes. The whole dataset is split into three sets: training, validation, and testing (denoted as train, val, and test) with 1464, 1449, and 1456 images, respectively. Following the common setting [4, 12], we also use the augmented training set provided by [8]. The resulting training set has 10582 weakly annotated images which we use to train our models. We compare our approach with other methods on both the val. and test sets and use mIoU as the evaluation metric.

**Implementation details.** We use VGG [27] pretrained from the ImageNet [5] 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% after 1 and 15 epochs respectively. For the weakly-supervised segmentation framework, following the setting of SEC [12], we use the DeepLab-CRFLargeFOV [4], which is a slightly modified version of the VGG network [27]. Implemented using Caffe [10], DeepLab-CRFLargeFOV [4] defines the input size as 321×321 and produces segmentation masks with size of 41×41. Our training procedure is the same as [12] at this stage. We run the SGD for 8000 iterations with the batch size of 15. The initial learning rate is $10^{-3}$ and it decreases by a factor of 10 for every 2000 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 [31], we separate them into two categories. For methods purely using image-level labels, we compare our GAIN-based SEC (denoted as GAIN in the table) with SEC [12], AE-PSL [31], TPL [11], STC [32] and 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$_{ext}$-based SEC (denoted as GAIN$_{ext}$ in the table) lies 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 [39], TransferNet [9] and MIL-seg [20] are included for comparison. 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 best performance with 55.3% and 56.8% in mIoU on these two sets, outperforming the SEC [12] baseline by 4.6% and 5.1%. Furthermore, GAIN outperforms AE-PSL [31] by 0.3% and 1.1%, and outperforms TPL [11] by 2.2% and 3.0%. These two methods are also proposed to cover more areas of the class of interest in attention maps. 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 [11, 31].

By implicitly using pixel-level supervision, our GAIN$_{ext}$-based SEC achieves 58.3% and 59.6% in mIoU when we use 200 randomly selected images with pixel-level labels (2% data of the whole dataset) as the extra supervision. It already performs 4% and 4.1% better than AF-MCG [39], which relies on the MCG generator [2] trained in a fully-supervised way on the PASCAL VOC. When the pixel-level supervision increases to 1464 images for our GAIN$_{ext}$, the performance jumps to 60.5% and 62.1%, which is a new state-of-the-art on this competitive benchmark for a challenging task. Figure 4 shows some qualitative results of semantic segmentation, indicating that GAIN-based methods help to discover more complete and accurate areas of classes of interest.

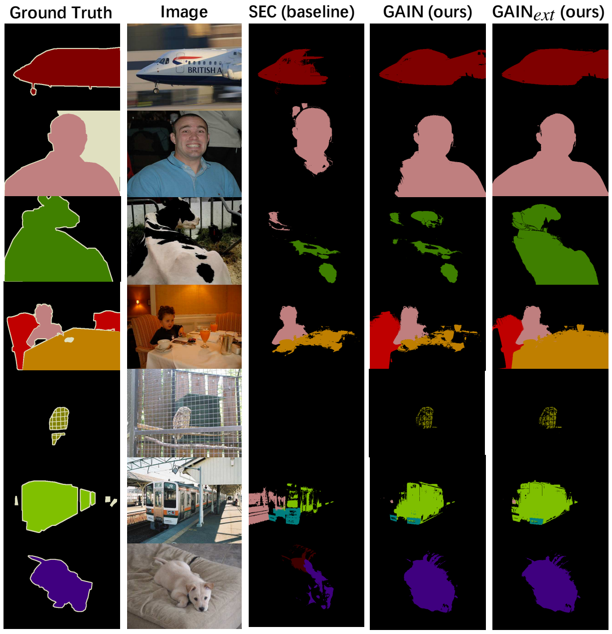

Figure 4. Qualitative results on PASCAL VOC 2012 *segmentation val.* set. They are generated by SEC (our baseline framework), our GAIN-based SEC and GAIN$_{ext}$-based SEC implicitly using 200 randomly selected (2%) extra supervision.

| Training Set | mIoU |
|---|---|
| 10K weak + 200 pixel | 58.3 |
| 10K weak + 400 pixel | 59.4 |
| 10K weak + 900 pixel | 60.2 |
| 10K weak + 1464 pixel | 60.5 |

Table 2. Results on PASCAL VOC 2012 *segmentation val.* set with our GAIN$_{ext}$-based SEC implicitly using different amount of pixel-level supervision for the attention map learning process.

We also show qualitative results of attention maps generated by GAIN-base methods in Figure 5, where GAIN covers more areas belonging to the class of interest compared with the Grad-CAM [24]. With only 2% of the pixel-level labels, the GAIN$_{ext}$ 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 Figure 5).

**More discussion of the GAIN$_{ext}$** We are interested in

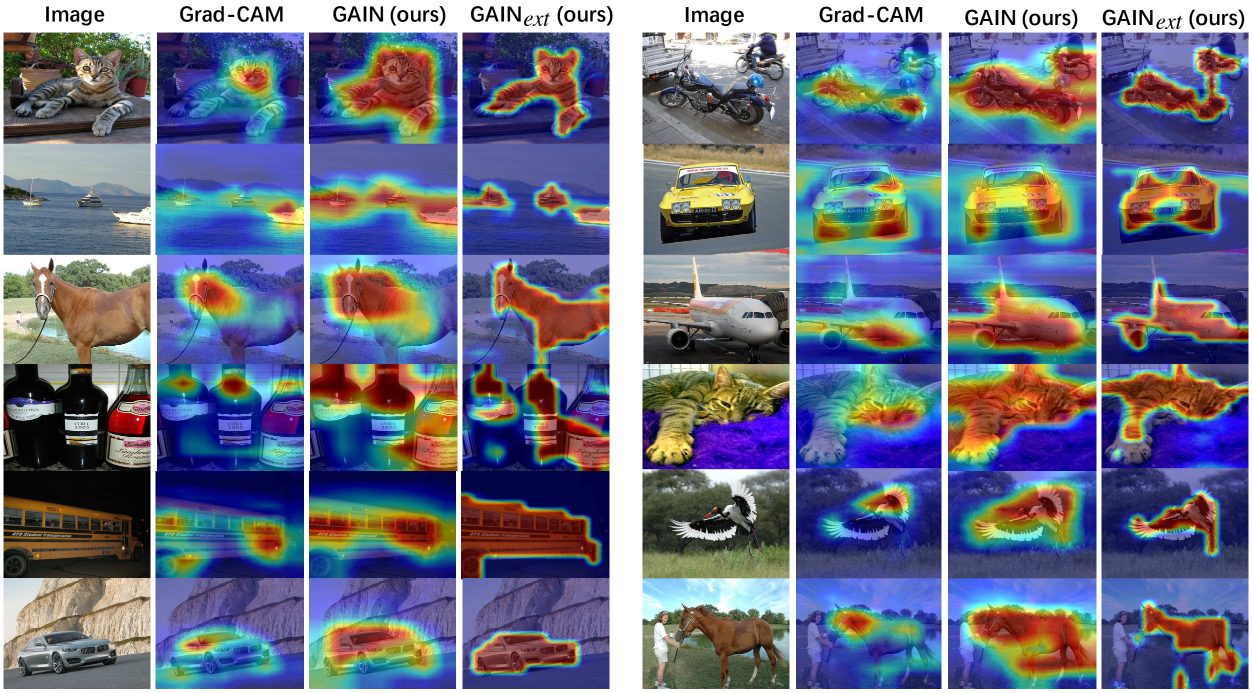

Figure 5. Qualitative results of attention maps generated by Grad-CAM [24], our GAIN and GAIN$_{ext}$ using 200 randomly selected (2%) extra supervision.

| Methods | Training Set | *val.* | *test* |
|---|---|---|---|
| SEC [12] w/o. CRF | 10K weak | 44.8 | 45.4 |
| GAIN w/o. CRF | 10K weak | 50.8 | 51.8 |
| GAIN$_{ext}$ w/o. CRF | 10K weak + 1464 pixel | 54.8 | 55.7 |

Table 3. Semantic segmentation results without CRF on PASCAL VOC 2012 *segmentation val.* and *test* sets. Numbers shown are mIoU.

finding out the influence of different amount of pixel-level labels on the performance. Following the same setting in Sec. 4.1, we add more randomly selected pixel-level labels to further improve attention maps and adopt them in the SEC [12]. From the results in Table 2, we find that the performance of the GAIN$_{ext}$ improves when more pixel-level labels are provided to train the network generating attention maps. Again, there are no pixel-level labels used to train the SEC segmentation framework.

We also evaluate performance on VOC 2012 *seg. val.* and *seg. test* datasets without CRF as shown in Table 3.

## 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.

**Boat experiment.** As shown in the Figure 1, the classification network trained on PASCAL VOC dataset focuses on sea and water regions instead of boats when predicting there are boats in an image. Therefore, the model failed to learn the right pattern or characteristics to recognize the boats, suffering from the bias in the training set. 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. Then we test the model trained without attention guidance, GAIN and GAIN$_{ext}$ described in Section 3.2 and 4.2 on this *Biased Boat* test dataset. Results are reported in Table 4. The models are exactly those trained in Sec 4.2. Some qualitative results are shown in Figure 6.

It can be seen that using GAIN with only image-level supervision, the overall accuracy on our *boat* dataset has been improved. 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 is still unable to fully decouple boat from water due to the biased training data.

On the other hand with GAIN$_{ext}$ training with small amount of pixel-level labels, similar levels of improvements are observed in both of the two scenarios. The reasons be-

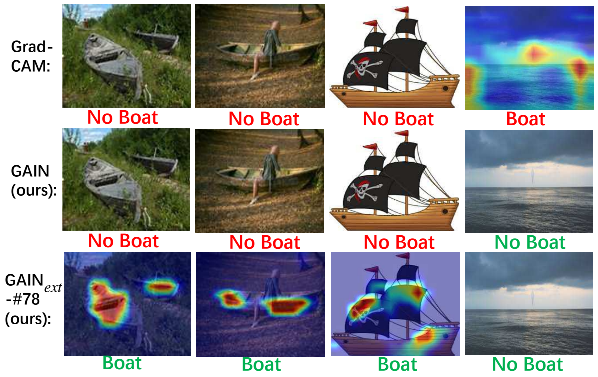

Figure 6. Qualitative results generated by Grad-CAM [24], our GAIN and GAIN$_{ext}$ on the *biased boat* dataset. **-#** denotes the number of pixel-level labels of *boat* used in the training which were randomly chosen from VOC 2012. Attention map corresponding to *boat* shown only when there are boats recognized.

| Test set | Grad-CAM | GAIN | GAIN$_{ext}$ (# 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% |

Table 4. Results comparison of Grad-CAM [24] with our GAIN and GAIN$_{ext}$ tested on the *biased boat* dataset for classification accuracy. **PL labels** denotes pixel-level labels of *boat* used in the training which are randomly chosen.

hind 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.

**Industrial camera experiment.** This one is designed for a challenging case to verify the model's generalization ability. We define two orientation categories for the industrial camera which is highly symmetric in shape. As shown in Figure 7, only features like gaps and small markers on the surface of the camera can be used to effectively distinguish their orientations. We then construct one training set and two test sets. *Training Set* and *Testing Set 1* are sampled from $\mathbf{D}_t$ without overlap. *Testing Set 2* is acquired with different camera viewpoints and backgrounds. There are 350 images for each orientation category in the *Training Set* resulting in 700 images in total and 100 images each in *Testing Set 1* and *Testing Set 2*. We train VGG-based Grad-CAM and our GAIN$_{ext}$ method on *Training Set*. In training GAIN$_{ext}$, manually drawn bounding boxes (20 for each classes taking up only 5% of the whole training data) on *critical areas* are used as external supervision.

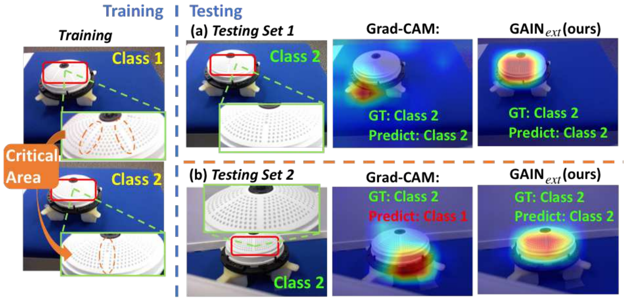

Figure 7. Datasets and qualitative results of our toy experiments. The critical areas are marked with red bounding boxes in each image. **GT** means ground truth orientation class label.

At testing procedure, though the Grad-CAM can well classify the images in the *Testing Set 1*, it only gets random guess results on *Testing Set 2* suffering from dataset bias. Instead, using GAIN$_{ext}$, the network is able to focus its attention on the area specified by the bounding box labels hence better generalization can be observed when testing with *Testing Set 2*. The results again suggest that our proposed GAIN$_{ext}$ has the potential of alleviating the impact of biases in training data, and guiding the learner to generalize better.

## 6. Conclusions

We propose a framework that provides direct guidance on the attention map generated by a weakly supervised learning deep neural network in order to teach the network to generate more accurate and complete attention maps. We achieve this by making the attention map not an afterthought, but a first-class citizen during training. Extensive experiments demonstrate that the resulting system confidently outperforms the state of the art without the need for recursive processing during run time. The proposed framework can be used to improve the robustness and generalization performance of networks during training with biased data, as well as the completeness of the attention map for better object localization and segmentation priors. In the future it may be illuminating to deploy our method on other high-level tasks than categorization and to explore for instance how a regression-type task may benefit from better attention.

## 7. 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, ONR Young Investigator Award N00014-14-1-0484 and U.S. Army Research Office Award W911NF-17-1-0367.

## References

[1] Pytorch. `http://pytorch.org/`.

[2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[3] 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 *CVPR*, 2015.

[4] 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*, 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[7] Y. Gong, S. Karanam, Z. Wu, K.-C. Peng, J. Ernst, and P. C. Doerschuk. Learning compositional visual concepts with mutual consistency. *arXiv preprint arXiv:1711.06148*, 2017.

[8] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[9] S. Hong, J. Oh, H. Lee, and B. Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016.

[10] 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 *MM*. ACM, 2014.

[11] D. Kim, D. Cho, D. Yoo, and I. So Kweon. Two-phase learning for weakly supervised object localization. In *ICCV*, 2017.

[12] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[13] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[14] M. Lin, Q. Chen, and S. Yan. Network in network. In *ICLR*, 2014.

[15] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[17] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[18] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015.

[19] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[20] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[21] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016.

[22] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*, 2017.

[23] 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 *ECCV*, 2016.

[24] 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 *ICCV*, 2017.

[25] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016.

[26] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[28] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

[29] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR Workshop*, 2015.

[30] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[31] 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 *CVPR*, 2017.

[32] 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 TPAMI*, 2017.

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[34] H. Zhang and V. M. Patel. Density-aware single image de-raining using a multi-stream dense network. *arXiv:1802.07412*, 2018.

[35] H. Zhang, V. Sindagi, and V. M. Patel. Image deraining using a conditional generative adversarial network. *arXiv:1701.05957*, 2017.

[36] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

[37] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.

[38] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *CVPR*, 2017.

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.