Attention Bridging Network for Knowledge Transfer

Kunpeng Li[1], Yulun Zhang[1], Kai Li[1], Yuanyuan Li[1] and Yun Fu[1,2]
[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA
[2]Khoury College of Computer Science, Northeastern University, Boston, MA

Abstract

The attention of a deep neural network obtained by back-propagating gradients can effectively explain the decision of the network. Therefore, attention maps can further be used to explicitly access to the network response to a specific pattern. Considering objects of the same category but from different domains share similar visual patterns, we propose to treat the network attention as a bridge to connect objects across domains. In this paper, we use knowledge from the source domain to guide the network's response to categories shared with the target domain. With weights sharing and domain adversary training, this knowledge can be successfully transferred by regularizing the network's response to the same category in the target domain. Specifically, we transfer the foreground prior from a simple single-label dataset to another complex multi-label dataset, leading to improvement of attention maps. Experiments about the weakly-supervised semantic segmentation task show the effectiveness of our method. Besides, we further explore and validate that the proposed method is able to improve the generalization ability of a classification network in domain adaptation and domain generalization settings.

1. Introduction

Since Convolutional Neural Networks (CNN) have achieved a lot of progress in many areas, various methods have been proposed recently to explain how they work [3, 37, 49]. Visual attention [35, 50] is one effective method to locate image regions that can contribute to the final prediction of the network. Attention maps can be obtained for a given input with back-propagation of the decision signal on a CNN [35]. They act as an effective way to analyze the network response and explain its decision.

Because of the close correlation with the network decision and response, attention maps can further be used to explicitly access to the network response to a specific pattern or category. Considering objects of the same category but from different domains share similar visual patterns, the network is likely to have similar responses to them. We are willing to explore the possibility of using network attention as a bridge to connect objects from different domains and transfer knowledge through it.

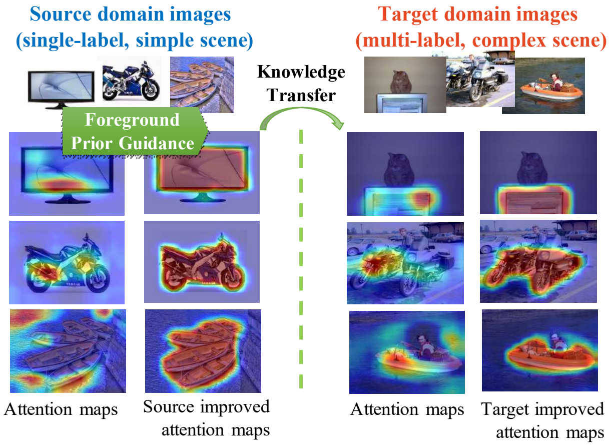

Figure 1. The proposed Attention Bridging Network (AttnBN) transfers the foreground prior from a simple single-label dataset (source domain) to another complex multi-label dataset (target domain), resulting in significant improvements of attention maps. By covering more complete regions of objects, these maps 1) help boost the performance of weakly-supervised semantic segmentation, and 2) guide the classification network to learn complete visual patterns of objects leading to better generalization ability.

Domains here could be datasets with different knowledge or priors. Transferring useful knowledge from one to the other could benefit the task of interest. Suppose there are two domains, the source and the target. Based on the understanding of network attention mechanism, we can use knowledge from the source domain to guide the network's response to categories shared with the target domain. With weights sharing and domain adversary training, this knowledge can be successfully transferred by regularizing the network's response to the same category in the target domain. We define this property as attention bridging mechanism and apply it in our model design.

We rely on two roles of network attention to design experiments accordingly to validate the effectiveness of attention bridging mechanism. (1) On the one hand, using image-level labels for training, attention maps of a classification network can provide localization information without extra labeling efforts. However, these attention maps often only cover most discriminative regions of target objects [15, 23, 39, 42, 45]. While these attention maps can still serve as reliable localization cues for tasks like weakly-supervised semantic segmentation [16], having integral attention maps that cover the target foreground objects completely have potential to further improve the performance. (2) On the other hand, network attention reflects the network's response and is related to the network's decision. integral attention, which covers complete regions of an object of interest, can guide the network to learn complete visual patterns of the object. This leads to the potential of boosting generalization ability of a classification network in both domain adaptation and domain generalization settings.

To this end, we propose Attention Bridging Network (AttnBN) for knowledge transfer across domains. As shown in Figure 1, taking weakly-supervised semantic segmentation as a task of interest, we aim to transfer useful information from a single-label dataset (simple source domain) to another multi-label dataset (complex target domain) to improve the attention maps. In the source domain, foreground-background prior, such as saliency information, can almost represent complete regions of objects in an image. However, this is not applicable for the target domain with multi-label images, whose foreground map may include multiple objects from different categories. Therefore, the foreground prior is regarded as advantageous knowledge in the source domain. AttnBN can transfer this knowledge across domains, resulting in a significant improvement of attention maps. By covering more complete regions of objects, these maps can act as better localization cues and help boost the performance of weakly-supervised semantic segmentation methods. Besides, they can also guide a classification network to learn more complete visual patterns of objects leading to better generalization ability.

To summarize, our contributions are: (1) We propose AttnBN that transfers knowledge across domains using network attention as a bridge. (2) Specifically, we transfer the saliency prior from a simple single-label dataset to another complex multi-label dataset to improve attention maps, so that these maps can cover the object holistically. (3) Experiments on PASCAL VOC benchmark [6] show that the improved attention maps can serve as better cues for weakly-supervised semantic segmentation models. (4) We also validate that AttnBN can improve the generalization ability of a classification network in both domain adaptation and domain generalization settings.

2. Related work

Network attention. Since Convolutional Neural Networks (CNN) have achieved great progress in many areas [20, 21, 48], a lot of methods have been proposed to analyze and explain deep neural networks [3, 25, 37, 49]. Based on them, visual attention is proposed to locate image regions that can contribute to the final prediction of the network. Inspired by a human visual attention model, [46] proposes a new back propagation method, Excitation Backprop, to hierarchy pass top-down signals downwards in the network. In [37, 40], error back-propagation based methods are proposed to visualize relevant regions for the activation of a hidden neuron or the network decision. CAM [50] shows that using an average pooling layer instead of fully-connected layers can help obtain attention maps which highlight task-related regions. Recently, CAM is extended by Grad-CAM [35] to various commonly used network architectures for tasks like visual question answering, image captioning and image classification. It generates reasonable visual explanations for various kinds of model decisions. Inspired by these methods that successfully model the response of the network, we explore the possibility of taking network attention as an interface to regularize the network's learning and response to a specific pattern.

Knowledge transfer. Transferring knowledge across datasets to benefit the task of interest has been widely studied in tasks of domain adaptation and transfer learning [22, 30]. Domain adaptation aims to solve the mismatch problem that data in different domains is sampled from different distributions. According to the specific application case, the transferred knowledge could be in the form of model parameters, feature representation or instances [30]. Different from these existing methods, we are trying to explore using attention of deep neural network as a bridge to transfer knowledge across domains. This is based on the hypothesis that objects of the same category but from different domains share similar visual patterns, therefore, the network is likely to have similar responses to them.

Weakly supervised methods. Weakly supervised learning [3, 37] aims to address the problem about labeled data scarcity and has recently attracted a lot of attention. Learning from image-level labels, attention maps of a classification network can provide localization information without extra labeling efforts for weakly-supervised semantic segmentation [2, 16, 23, 24, 45], object localization [47, 50], object detection [44] and etc. However, the attention map of a trained classification network only cover most discriminative regions of target objects, which is not good enough for these tasks that aim to localize complete, interior and dense regions. To reduce this gap, [39] proposes to randomly hide patches in each training image, so that the network would be forced to discover other relevant regions when the discriminative parts are missing. It can be treated as a useful data augmentation method. However, it relies on a strong assumption that foreground objects would not be completely hidden by patches. More recently, [23, 42, 47] use the

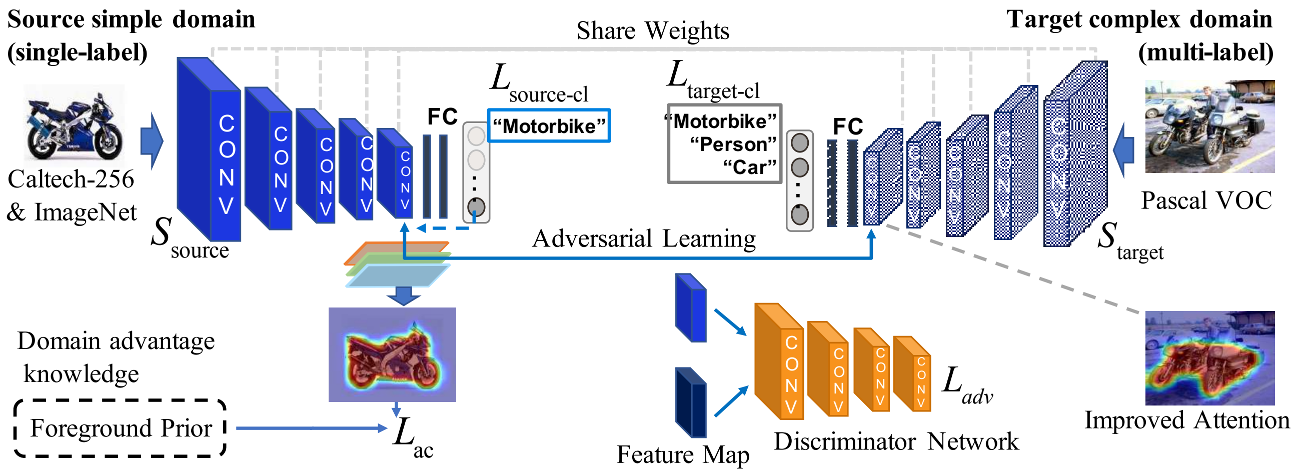

Figure 2. AttnBN includes one discriminator network and two streams of classification networks. Attention map are end-to-end trainable and jointly optimized by four loss functions. Advantageous knowledge (saliency prior here) from the source domain guides the network's response to categories shared with the target domain. With weights sharing and domain adversary training, this advantageous knowledge can be successfully transferred by regularizing the network's response to the same category in the target domain.

adversarial erasing strategy to guide the attention maps to cover more complete foreground objects. In [45], dilated convolutional blocks with various dilation rates are added to a classification network. Experiments validate that different dilation rates could help transfer the surrounding discriminative information to non-discriminative object parts. Different from these approaches, we explore the attention bridging mechanism to transfer knowledge across domains to get more complete attention maps, which can benefit the semantic segmentation task.

3. Attention Bridging Network

Foreground-background priors as well as the scene complexity are unequal for the single-label dataset and the multi-label dataset. The foreground part of a single-label image can almost represent complete regions for the particular class. While a multi-label image may include multiple objects from different classes in its foreground. Therefore, transferring this advantageous knowledge across domains is strongly motivated. In this section, we describe our Attention Bridging Network (AttnBN) to achieve this goal. Network attention acts as a bridge to connect different domains.

Overview of the proposed model. Suppose we have two datasets, D_s including images of single-label as the source domain and D_m for images with multi-label as the target domain. They are composed of K_s and K_m images from N classes respectively. We aim to transfer knowledge about foreground-background priors and boundary constrains from the source domain to the target domain. As shown in Figure 2, our AttnBN includes one discriminator network and two streams of classification networks, Source Stream S_{source} for the source domain and Target Stream S_{target} for the target domain, which share parameters with each other. Domain advantage knowledge in D_s (saliency map here) is used to directly guide Stream S_{source} to focus on more complete regions of salient foregrounds when learning to recognize classes. This will simultaneously regularize the Stream S_{target}'s response to the same class in the target domain during the training process benefit from the weight sharing and attention mechanism. Besides, since network attention is a reflection of the network response, it is closely related to the learned feature space. Therefore, we integrate the adversarial learning scheme to encourage the network to learn domain-invariant features, which has potential to boost the effect of attention bridging.

Obtain trainable attention maps. To make sure the guidance from saliency maps can directly regularize the network response to both domains, we first generate trainable attention following [23, 35, 47]. Specifically, in stream S_{source}, for a input image I, F_i represents the activation of feature map i in the last convolutional layer whose features have better trade-off between detailed spatial information and high-level semantics [37]. Class specific attention maps can be obtained by computing the gradient of the score y^c for class c, with respect to activation maps $F_i(x, y)$. A global average pooling operation is then performed on these gradients [26] to get importance weights w_i^c for neurons as follows,

$$w_i^c = \frac{1}{H} \sum_{x,y} \frac{\partial y^c}{\partial F_i(x,y)}, \tag{1}$$

where H is the size of the convolutional feature map F_i (equals to 196, 14×14 in the case of VGG [38]).

Based on recent work [50], each unit F_i is expected to be activated by specific visual patterns within its receptive field. Therefore, as shown in Eq. 2, the class attention map M^c is a weighted w_i^c sum of these visual patterns presence at different locations F_i followed by a ReLU operation. This is equivalent to treating weight matrix w^c as a kernel and doing a 2D convolution operation over feature maps F_i [23]:

$$M^c = \sigma(\sum_i w_i^c F_i) = \sigma\left(\mathrm{Conv}\left(F, w^c\right)\right), \tag{2}$$

where $\sigma(\cdot)$ represents the ReLU operation.

Attention bridging. In the source domain, we have saliency maps A for these single-label images to provide knowledge about foreground-background priors and boundary constrains. A could constrain the network attention learning to encourage it to focus on more complete regions of salient foregrounds when recognizing classes. As shown in Eq. 3, L2 loss is adopted to calculate the attention constrain loss L_{ac} for stream S_{source} to achieve this goal.

$$L_{ac} = \left(M^c - A\right)^2, \tag{3}$$

where A is the saliency map for a given image, M^c is the attention map towards its single-label ground-truth class c.

For the classification component in S_{source}, the single-label of the source domain image is converted to one-hot vector $l = \{l_1, l_2, ..., l_N\}$, where N is the number of ground truth classes. Then, a multi-label soft margin loss is used here as L_{s-cl} to make sure regions within the network attention will help to recognize classes. We use the same loss denoted as L_{t-cl} for target domain stream S_{target}.

$$L_{s-cl}(o, l) = -\sum_j l_j \log(p_j) + (1 - l_j)\log(1 - p_j), \tag{4}$$

where $p_j = (1 + e^{-o_j})^{-1}$, o_j is the output of last fully-connected layer for the classification component of S_{source}.

For the adversarial learning part, the training objective is to learn domain-invariant features, which can boost the effect of attention bridging. Since network attention is closely related to the network response especially the feature map of the last convolutional layer F in our current implementation, we forward the F_s of stream S_{source} and F_t of stream S_{target} to a fully-convolutional discriminator D. Then a cross-entropy loss L_d for the two classes (source and target) is adopted to train D.

$$L_d = -(1 - d)\log(D(F_s)) - d\log(D(F_t)), \tag{5}$$

where $d = 0$ if the sample comes from the target domain and $d = 1$ if it is from the source domain.

Then, when training the classification network, for the samples I_t from the target domain, we forward the feature map of the last convolutional layer F_t of stream S_{target} to the discriminator and use following adversarial loss to help learn domain-invariant features by fooling the discriminator network:

$$L_{adv} = -\log(D(F_t)). \tag{6}$$

Our final attention bridging loss L_{ab} is the weighted sum of the classification loss L_{s-cl}, L_{t-cl} and attention constrain loss L_{ac} as defined in Eq. 7.

$$L_{ab} = L_{s-cl} + \lambda_1 L_{ac} + L_{t-cl} + \lambda_{adv} L_{adv}, \tag{7}$$

where L_{t-cl} is the classification loss for target domain stream S_{target} which uses the same function with L_{s-cl}. Hyper-parameters λ_1 and λ_{adv} balance the four losses. We set $\lambda_1 = 2$ and $\lambda_{adv} = 10^{-3}$ in all of our experiments.

Based on weights sharing and attention mechanism, L_{ab} can transfer knowledge from the source domain to the target domain to improve attention maps.

4. Experiments for semantic segmentation

The proposed AttnBN transfers knowledge across domains to improve attention maps, so that these maps can cover more complete object of interest. To verify this, we take the semantic segmentation as the task of interest to validate the effectiveness of AttnBN. (1) We first conduct ablation studies to incrementally validate each component of AttnBN (Eq. 7). To directly evaluate attention maps of each ablation model, we combine attention maps of different classes as semantic segmentation results (Section 4.2). (2) We also take attention maps as localization cues to train weakly-supervised semantic segmentation models and generate results for further evaluation (Section 4.3).

4.1. Experimental setup

Datasets. We use PASCAL VOC 2012 segmentation dataset as the target domain dataset which includes multi-label images of 20 categories. The images are 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 [5, 16], we use the augmented training set provided by [9], which leads to 10582 weakly annotated images for the training set of the target domain. Then, subsets of Caltech-256 [8] and ImageNet CLS-LOC [34] within these 20 VOC categories are combined together, resulting in around 20K single-label images as the source domain dataset. We train our model using images in both source and target domain with only image-level class labels and evaluate it on PASCAL VOC 2012 segmentation benchmark val. and test sets. The standard mean intersection-over-union (mIoU) metric is used to report quantitative results.

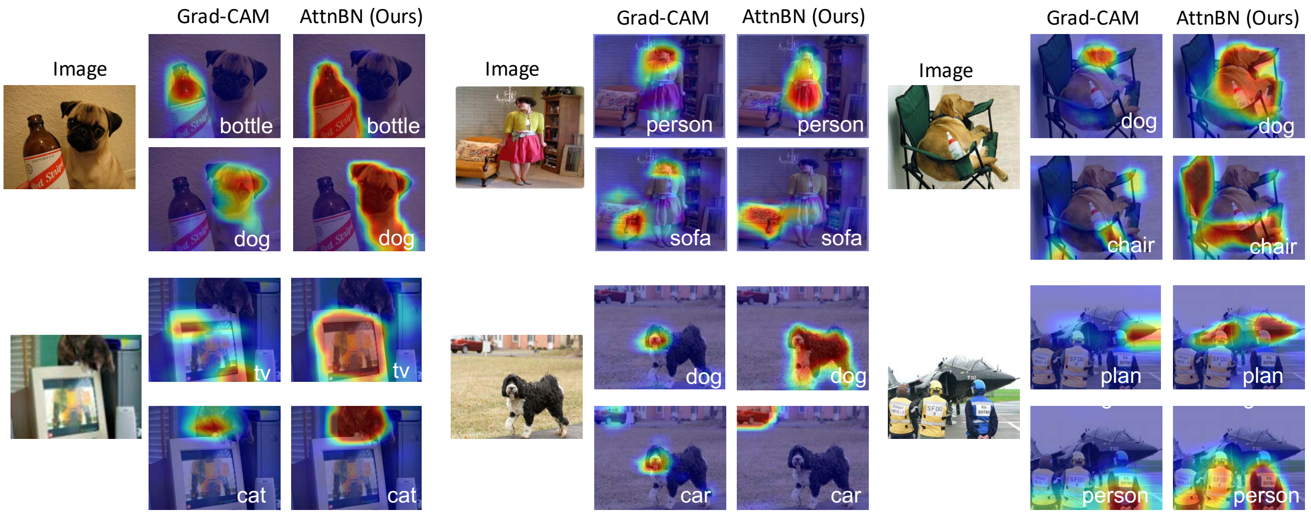

Figure 3. Qualitative results of attention maps. AttnBN focuses on more complete regions belonging to the class of interest than the baseline model Grad-CAM [35].

Methods	b.g.	plane	bike	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	hors.	moto.	pers.	plant	sheep	sofa	train	tv	mIoU
Results on the val. set:																						
L_{t-cl} (Grad-CAM) [35]	74.0	34.2	19.5	33.1	18.6	25.0	41.7	27.9	36.1	11.4	26.3	20.7	30.4	29.0	41.5	40.2	21.6	32.8	18.2	23.6	35.3	30.2
$L_{t-cl} + L_{s-cl}$	74.9	38.1	20.3	34.2	21.1	26.8	38.5	31.6	34.9	10.0	31.7	25.5	29.1	30.7	41.0	41.2	21.9	32.5	19.7	23.5	37.5	31.7
$L_{s-cl} + L_{ac}$	78.2	57.2	24.1	48.2	36.3	37.2	48.5	45.3	37.2	10.8	34.6	13.7	34.8	35.0	41.4	34.9	21.5	38.7	20.4	39.2	31.7	37.5
$L_{s-cl} + L_{ac}, L_{t-cl}$	80.5	60.9	26.5	47.5	37.9	37.8	51.3	46.1	36.9	11.1	34.4	13.3	37.1	37.7	43.8	36.5	19.9	40.2	19.6	41.3	31.0	39.4
$L_{s-cl} + L_{ac} + L_{t-cl}$	82.8	64.4	26.8	59.7	44.0	48.5	65.1	56.3	58.8	10.3	53.8	17.2	59.6	50.3	49.5	54.9	27.3	60.1	25.4	56.7	38.6	46.2
AttnBN	83.7	68.6	25.4	62.3	47.4	52.4	66.9	61.7	63.2	10.5	57.3	18.7	62.8	53.1	52.3	58.2	30.2	63.9	28.1	60.7	42.2	**50.9**
Results on the test set:																						
L_{t-cl} (Grad-CAM) [35]	76.2	36.6	20.0	32.3	15.6	30.8	39.3	26.1	37.6	12.3	25.0	27.7	30.3	30.6	43.8	41.2	24.7	35.7	23.4	19.8	38.5	31.7
$L_{t-cl} + L_{s-cl}$	76.9	36.8	19.9	34.8	15.0	29.2	38.1	28.0	35.8	12.7	32.2	30.5	29.6	31.0	38.4	41.3	27.6	35.2	25.0	18.9	38.3	32.2
$L_{s-cl} + L_{ac}$	78.7	58.5	24.7	49.4	30.8	31.6	49.3	47.1	37.7	11.3	35.8	14.3	36.4	35.6	42.8	36.2	20.7	39.8	21.9	44.7	32.4	38.7
$L_{s-cl} + L_{ac}, L_{t-cl}$	82.3	60.0	25.5	52.4	33.7	34.7	50.6	46.8	39.7	10.3	36.1	16.4	40.7	34.5	43.1	38.6	21.0	40.7	20.8	47.0	31.1	39.8
$L_{s-cl} + L_{ac} + L_{t-cl}$	83.4	60.2	27.5	60.8	35.2	49.2	65.7	56.9	59.2	11.0	50.2	17.8	59.2	51.5	52.1	55.4	29.8	57.8	29.4	57.2	38.9	47.5
AttnBN	84.7	63.3	26.8	63.0	39.2	51.6	68.9	60.9	62.1	13.0	52.7	18.8	62.2	59.3	56.4	57.7	34.3	59.5	34.5	62.4	43.4	**51.2**

Table 1. Ablation studies on VOC 2012 *segmentation val.* and *test* sets. We directly evaluate attention maps of each ablation model. This is achieved by combining attention maps of predicted classes as semantic segmentation results for evaluation (Section 4.2).

Implementation details. We use VGG [38] pre-trained from the ImageNet [34] as the backbone classification network for our AttnBN. For the discriminator network, we adopt a similar architecture from [33] but use all fully-convolutional layers to retain the spatial information. An up-sampling layer with a scale factor of 5 is used to rescale the last convolutional layer features from both domains as the input to the discriminator. We use $\lambda_1 = 2$ and $\lambda_{adv} = 10^{-3}$ in Eq. 7 in all of our experiments and jointly train the discriminator with the AttnBN network. Saliency maps for the source domain dataset are obtained by using the method and trained model provided by [27]. We use Pytorch [1] to implement our model and use the stochastic gradient descent (SGD) to train it for 30 epochs. We start training with learning rate of 10^{-4} for 20 epochs, and then lower the learning rate to 10^{-5} for the rest 10 epochs.

4.2. Ablation studies with direct evaluation

Quantitative results. To directly evaluate attention map quality, we combine attention maps of classes which are predicted by the trained model as semantic segmentation maps. When there are overlaps between attention maps of different classes in a single image, we choose the one with the largest prediction score. No post-processing is used. Better segmentation results are expected to be obtained when complete and accurate attention maps are combined. Ablation studies on PASCAL VOC 2012 *segmentation val.* set and *segmentation test.* set are shown in Table 1. We start from the baseline model only using classification losses L_{t-cl} and the target domain data for training, which is actually Grad-CAM [35] model. It achieves mIoU of 30.2 on val. set and 31.7 on test set. We then add classification losses L_{s-cl} and source data for training. The improvement is around (around 1% of mIoU. This shows that a mere in-

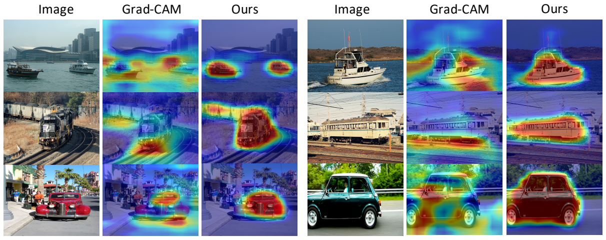

Figure 4. Qualitative results of attention maps obtained by the baseline model Grad-CAM [35] and AttnBN. AttnBN can guide the network focus less on the background contents that always come together with the objects and are helpful for the recognition, like water with boats, the rail with trains and road with cars.

Methods	Training Set	val.	test
AttnBN	10K (T) + 2K (S)	45.1	45.7
AttnBN	10K (T) + 10K (S)	49.2	50.3
AttnBN	10K (T) + 20K (S)	**50.9**	**51.2**

Table 2. Direct evaluation of AttnBN on Pascal VOC 2012 dataset with different amount of source domain data available. **T** and **S** denote target and source domain respectively. Numbers are mIoU.

Method	Supervision	Sal.	val.	test
MIL-seg$_{(\mathrm{CVPR}'15)}$ [31]	700K W	✗	40.6	42.0
SEC$_{(\mathrm{ECCV}'16)}$ [16]	10K W	✗	50.7	51.7
STC$_{(\mathrm{PAMI}'16)}$ [43]	50K W	✓	49.8	51.2
TransNet$_{(\mathrm{CVPR}'16)}$ [10]	10K W+60K P	✗	52.1	51.2
AF-MCG$_{(\mathrm{ECCV}'16)}$ [32]	10K W+1.4K P	✗	54.3	55.5
TPL$_{(\mathrm{ICCV}'17)}$ [15]	10K W	✗	53.1	53.8
AE-PSL$_{(\mathrm{CVPR}'17)}$ [42]	10K W	✓	55.0	55.7
Oh et al.$_{(\mathrm{CVPR}'17)}$ [29]	10K W	✓	55.7	56.7
CrawlSeg$_{(\mathrm{CVPR}'17)}$ [11]	970K W	✗	58.1	58.7
WebS-i2$_{(\mathrm{CVPR}'17)}$ [14]	19K W	✗	53.4	55.3
DCSP$_{(\mathrm{BMVC}'17)}$ [4]	10K W	✓	58.6	59.2
MEFF$_{(\mathrm{CVPR}'18)}$ [7]	10K W	✗	-	55.6
AffinityNet$_{(\mathrm{CVPR}'18)}$ [2]	10K W	✗	58.4	60.5
Shen et al.$_{(\mathrm{CVPR}'18)}$ [36]	86.7K W	✗	58.8	60.2
DilConv$_{(\mathrm{CVPR}'18)}$ [45]	10K W	✓	60.4	60.8
GAIN$_{(\mathrm{CVPR}'18)}$ [23]	10K W	✓	55.3	56.8
MCOF$_{(\mathrm{CVPR}'18)}$ [41]	10K W	✓	56.2	57.6
DSRG$_{(\mathrm{CVPR}'18)}$ [13]	10K W	✓	59.0	60.4
AttnBN (ours)	12K W	✓	61.7	62.3
AttnBN (ours)	30K W	✓	**62.1**	**63.0**

Table 3. Comparison with state-of-the-art weakly-supervised semantic segmentation methods on Pascal VOC 2012 dataset. "W" means weak supervision from image-level labels and "P" means strong supervision from pixel-level labels. "Sal." represents using saliency prior. Results shown are based on VGG backbone.

creasing of data from the source domain only is of trivial benefit. We also test the model trained with $L_{s-cl} + L_{ac}$ on the source domain. It achieves much better improvement, which benefits from the guided attention learning mechanism in our model. However, the improvement is still not good enough due to the domain shift between the source and target. Then we take use of the target domain data and perform ether fine-tuning (noted as $L_{s-cl} + L_{ac}$, L_{t-cl}) or two-domain joint training (noted as $L_{s-cl} + L_{ac} + L_{t-cl}$). Results validate the importance of joint training and weight sharing strategies in attention bridging mechanism. Finally we add adversary training L_{adv} (noted as AttnBN in the table) to further help learn domain-invariant features and boost the effect of attention bridging. AttnBN achieves mIoU of 50.9 on val. set and 51.2 on test set, which shows a huge improvement upon the baseline model Grad-CAM by successfully transferring knowledge across domains.

Qualitative results. As shown in Figure 3, AttnBN focuses on more complete regions belonging to the class of interest than the baseline model Grad-CAM [35]. This mainly because AttnBN learns concepts of integral objects from the images with simple scenes in the source domain, and successfully transfers this knowledge to the target domain, where images include complex scenes.

Analysis of classes with huge improvements. We further analyze the detailed quantitative results for each class and get some interesting findings. For classes with huge improvements like *boat* (28.8% for *val.*), *train* (37.1%) and *car* (33.8%). We show several qualitative results in Figure 4 and find that Grad-CAM would focus on background contents when predicting classes. That is because these background contents always come together with the objects and are helpful for the recognition, like water with boats, the rail with trains and road with cars. With these bias information within the dataset, only constrained by classification loss will make the network consider these background contents as one of the most prominent feature characterizing the classes. Our AttnBN can well handle this problem by transferring knowledge of foreground prior from the source domain to guide the network to learn the correct concept.

Using different amount of source domain data. We are also interested in finding out the influence of using different amount of source domain data. Therefore, we randomly sample from the source domain dataset with a ratio of 0.1 and 0.5 to obtain two new source domain subsets. We train AttnBN on these two subsets separately. Following directly evaluation mentioned before, we still combine attention maps as semantic segmentation results and do evaluations. Quantitative results on PASCAL VOC 2012 are shown in Table 2. We find that the performance of AttnBN improves when more source domain data is available. Besides, only using 10% source domain data with about 2K images can already improve upon Grad-CAM by 14.9% and 14.0% of mIoU on the val. and test sets. This shows the effectiveness of AttnBN for transferring advantageous knowledge (foreground prior here) across domains.

4.3. Act as priors for weakly-supervised framework

In the weakly-supervised setting, recent methods [13, 15, 16, 42, 45] mainly rely on localization cues obtained by Grad-Cam [35] or CAM[50], and consider other con-

Methods	b.g.	plane	bike	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	hors.	moto.	pers.	plant	sheep	sofa	train	tv	mIoU
SEC [16]	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
TransferNet [10]	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
AE-PSL [42]	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
DilConv [45]	89.5	85.6	34.6	75.8	61.9	65.8	67.1	73.3	80.2	15.1	69.9	8.1	75.0	68.4	70.9	71.5	32.6	74.9	24.8	73.2	50.8	60.4
GAIN [23]	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
DSRG [13]	87.5	73.1	28.4	75.4	39.5	54.5	78.2	71.3	80.6	25.0	63.3	25.4	77.8	65.4	65.2	72.8	41.2	74.3	34.1	52.1	53.0	59.0
AttnBN	89.5	82.0	30.1	76.2	57.9	65.3	80.7	75.6	79.5	16.8	68.9	19.7	76.4	70.4	67.7	71.8	40.1	72.1	37.2	73.1	53.7	**62.1**

Table 4. Detailed results of state-of-the-art weakly supervised semantic segmentation methods on 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
SEC [16]	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
TransferNet [10]	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
AE-PSL [42]	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
DilConv [45]	89.8	78.4	36.2	82.1	52.4	61.7	64.2	73.5	78.4	14.7	70.3	11.9	75.3	74.2	81.0	72.6	38.8	76.7	24.6	70.7	50.3	60.8
GAIN [23]	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
DSRG [13]	87.9	69.5	32.1	74.2	33.7	59.4	74.9	71.5	80.1	21.9	66.8	32.7	76.4	72.5	76.6	73.4	49.9	73.8	43.4	42.0	55.2	60.4
AttnBN	89.9	75.7	32.9	73.5	49.9	60.4	78.1	76.5	77.4	19.9	72.0	27.4	73.8	72.7	77.2	72.3	51.2	77.3	37.9	73.5	53.6	**63.0**

Table 5. Detailed results of state-of-the-art weakly supervised semantic segmentation methods on VOC 2012 *segmentation test* set.

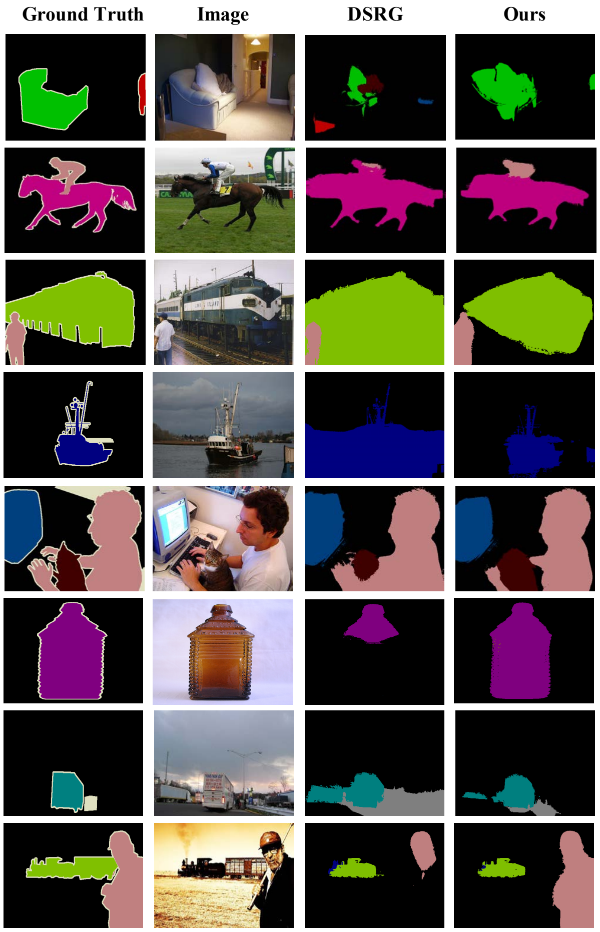

Figure 5. Qualitative weakly-supervised semantic segmentation results of DSRG and our method.

straints like object boundaries to train a segmentation network. The performance of these methods is highly influenced by the quality of localization cues. Compared with attention maps obtained by Grad-Cam and CAM that only cover small and the most discriminative regions, attention maps of AttnBN can locate more complete regions belonging to the class of interest. Therefore, they have potential to help improve the performance of weakly-supervised segmentation methods. To validate this, we take attention maps from AttnBN as foreground localization cues for the existing weakly-supervised semantic segmentation method DSRG [13] and use [12] to obtain background cues. We then train DSRG with VGG as the backbone to generate segmentation results using the same inference procedure, as well as parameters of CRF [17].

We report quantitative results on Pascal VOC 2012 dataset in Table 3. Our results are noted as AttnBN. We make extensive comparisons with state-of-the-art weakly-supervised semantic segmentation solutions with different configurations. From the results, we can find AttnBN obtains the best performance with 62.1% and 63.0% in mIoU on val. and test sets respectively. Compared with baseline model DSRG, AttnBN provides a performance gain with 3.1% on val. set and 2.6% on test set. Note that our training of semantic segmentation network follows the same setting and training data (only PASCAL VOC 2012) with DSRG as well as other recent works. Different amount of source data is only used when training the attention map generation model (AttnBN). Consider ablation studies in Table 1, results of $L_{t-cl} + L_{s-cl}$ show that a mere increasing of data from the source domain only is of trivial benefit. The main improvement is due to effectiveness of knowledge transfer and better attention maps. This verifies that AttnBN can generate high quality attention maps as cues to improve the performance of weakly-supervised methods.

Besides, comparing with methods also focusing on knowledge transfer, such as Shen et al. [36], CrawlSeg [11], WebS-i2 [14], TransNet [10], our methods achieve better

Methods	Setting	mAP on Target test
L_{s-cl}	Generalization	61.3
$L_{s-cl} + L_{ac}$	Generalization	66.7
$L_{s-cl} + L_{adv}$	Adaptation	64.5
$L_{s-cl} + L_{ac} + L_{adv}$	Adaptation	71.0

Table 6. Quantitative results for classification in unsupervised domain adaptation and domain generalization settings between single-label domain and multi-label domain. “Target test” represents the VOC 2012 testing set.

Methods						
Source:	C	I	V	I	V	C
Target:	I	C	I	V	C	V
Training setting: Generalization						
L_{s-cl}	0.77	0.98	0.87	0.83	0.97	0.62
$L_{s-cl} + L_{ac}$	0.83	0.99	0.90	0.88	0.98	0.71
Training setting: Adaptation						
$L_{s-cl} + L_{adv}$	0.81	0.98	0.88	0.85	0.98	0.67
$L_{s-cl} + L_{ac} + L_{adv}$	0.87	0.99	0.91	0.89	0.99	0.74

Table 7. Quantitative results for classification in domain adaptation and domain generalization settings between PASCAL VOC2007, Caltech-101 (C), and ImageNet (I). Numbers shown are accuracy.

performance using less extra data. Furthermore, AttnBN outperforms AE-PSL [42] by 7.1% and 7.3%, DilConv [45] by 1.7% and 2.2%, GAIN [23] by 6.8% and 6.2% on val and test set respectively. These methods are also proposed to generate better attention maps and they also take use of saliency priors.

Table 4 and Table 5 show detail results of each class on PASCAL VOC 2012 *segmentation val.* set and *segmentation test.* set. Figure 5 shows qualitative results of semantic segmentation obtained by DSRG and AttnBN. From the first three rows, we can find that AttnBN can help to generate better segmentation results based on more complete attention maps. From the last two rows, our results include less background than DSRG. It is mainly because AttnBN can guide the network focus less on the background contents that always come together with the objects and are helpful for the recognition, like water with boats, the rail with trains. This is consistent with the analysis in Figure 4.

5. Experiments for domain adaptation and domain generalization

In addition to serving as localization cues for weakly-supervised method, attention maps also reflect the network responses and related to network's predication. Integrate attention can help to learn the complete visual pattern of objects and has potential to boost generalization performance of the network. Therefore, we further validate this by designing experiments for the classification task in the domain adaptation and domain generalization settings.

For the domain generalization setting, we treat one of the dataset as the source domain and the other two unseen datasets as unseen domains. Only data and label in the source domain are available during training. For the domain adaption setting, we take one of the dataset as the source domain and one of the other two datasets as the target domain. Only data and labels in the source domain as well as data in the target domain is available during training, no label in the target domain is used.

We first do experiments using the two domains as described in Section 4.1 to explore the adaptation and domain generalization from the single-label domain to the multi-label domain. The baseline model here is VGG [38] trained with classification loss L_{s-cl} defined in Eq. 7. We test our model trained with $L_{s-cl}$and attention constrain loss L_{ac} for domain generalization setting. We also report results of the models trained with $L_{s-cl} + L_{ac} + L_{adv}$ or with $L_{s-cl} + L_{adv}$ for the domain adaptation setting, where L_{adv} is the adversary training loss. For all these four models, no labels in the target domain are used during training. From Table 6, we find AttnBN can help learn domain invariant features benefiting from an integral attention.

We further validate the effectiveness of AttnBN in more general cases. Following [18, 19, 28], we use images of 5 common object categories (bird, car, chair, dog, and person) of the PASCAL VOC2007 (V) [6], Caltech-101 (C) [8], and ImageNet (I) [34] datasets to design experiments in two settings. We test the same four models defined in our last experiment and report results in Table 7. We can find AttnBN helps to improve the generalization ability of the classification network. This validates the advantage of the integral attention and the strength of attention bridging mechanism.

6. Conclusion

We propose AttnBN that can transfer knowledge across domains using network attention as a bridge. This is based on our understanding that network attention can be used to explicitly access to the network response to objects of the same category but from different domains. Experiments for weakly-supervised semantic segmentation demonstrate the effectiveness of the proposed method. We also validate that our method can help improve the generalization ability of a classification network in both domain adaptation and domain generalization settings. In the future, since the source domain is quite simple in our case, we will try to use unsupervised or weakly-supervised saliency detection methods to generate foreground prior for the source domain. We will also explore more knowledge transfer scenarios that are related to the network attention.

7. Acknowledgments

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] Pytorch. http://pytorch.org/.

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018.

[3] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *CVPR*, 2015.

[4] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *BMVC*, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

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

[7] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*, 2018.

[8] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[9] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[10] Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016.

[11] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017.

[12] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

[13] X. Huang, Z.and Wang, J. Wang, W. Liu, and J Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018.

[14] Bin Jin, Maria V Ortiz Segovia, and Sabine Süsstrunk. Webly supervised semantic segmentation. In *CVPR*, 2017.

[15] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *ICCV*, 2017.

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

[17] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[18] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*. IEEE, 2017.

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[20] Kai Li, Zhengming Ding, Kunpeng Li, Yulun Zhang, and Yun Fu. Support neighbor loss for person re-identification. In *ACM Multimedia*, 2018.

[21] Kunpeng Li, Yu Kong, and Yun Fu. Multi-stream deep similarity learning networks for visual tracking. In *IJCAI*, 2017.

[22] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, 2019.

[23] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018.

[24] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *TPAMI*, 2019.

[25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019.

[26] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICLR*, 2014.

[27] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.

[28] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*. IEEE, 2017.

[29] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, 2017.

[30] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2010.

[31] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[32] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016.

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[36] T. Shen, G. Lin, C. Shen, and I. Reid. Bootstrapping the performance of webly supervised semantic segmentation. In *CVPR*, 2018.

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[39] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

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

[41] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018.

[42] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017.

[43] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 2016.

[44] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *ECCV*, 2018.

[45] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, 2018.

[46] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

[47] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018.

[48] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019.

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2014.

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.