L2G: A Simple Local-to-Global Knowledge Transfer Framework for Weakly Supervised Semantic Segmentation

Peng-Tao Jiang[1]   Yuqi Yang[1]   Qibin Hou[1*]   Yunchao Wei[2]
[1]TMCC, CS, Nankai University   [2]Beijing Jiaotong University
pt.jiang@mail.nankai.edu.cn andrewhoux@gmail.com

Abstract

Mining precise class-aware attention maps, a.k.a, class activation maps, is essential for weakly supervised semantic segmentation. In this paper, we present L2G, a simple online local-to-global knowledge transfer framework for high-quality object attention mining. We observe that classification models can discover object regions with more details when replacing the input image with its local patches. Taking this into account, we first leverage a local classification network to extract attentions from multiple local patches randomly cropped from the input image. Then, we utilize a global network to learn complementary attention knowledge across multiple local attention maps online. Our framework conducts the global network to learn the captured rich object detail knowledge from a global view and thereby produces high-quality attention maps that can be directly used as pseudo annotations for semantic segmentation networks. Experiments show that our method attains 72.1% and 44.2% mIoU scores on the validation set of PASCAL VOC 2012 and MS COCO 2014, respectively, setting new state-of-the-art records. Code is available at https://github.com/PengtaoJiang/L2G.

1. Introduction

Deep learning algorithms [37,41,64] have promoted the rapid development of the semantic segmentation task in recent years. However, training a deep neural network for semantic segmentation requires a large number of pixel-wise accurate labels, which consume lots of human labors and resources. Recently, to reduce the reliance on accurate annotations, researchers have attempted to study semantic segmentation based on cheap supervisions, such as bounding boxes [12,42], scribbles [36,50], points [4], and image-level labels [23, 53]. Among these weak supervisions, image-level labels only provide information on the existence of the target object categories, making them more popular than

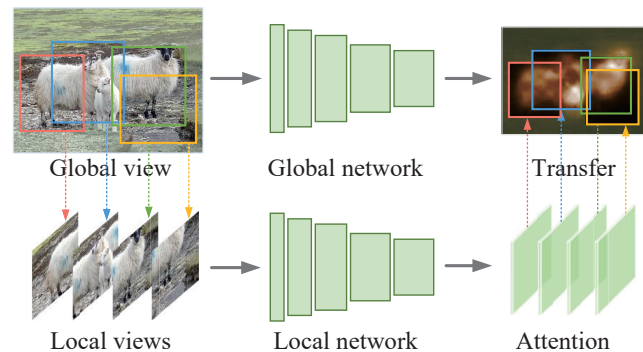---

*Qibin Hou is the corresponding author.

Figure 1. Conceptual working pipeline of the proposed method. We utilize the attention maps for local views with rich details extracted from the local network to teach the global network. This enables the global network to learn the rich local details knowledge from the local network online and thereby more integral object attentions.

other supervisions due to the easy way to collect. In this paper, we also focus on weakly supervised semantic segmentation (WSSS) based on image-level labels.

Speaking of WSSS, one of the most important components should be the class activation map (CAM) [65] which contains both semantic and location information about the target objects and can be used as pseudo pixel-level annotations for training segmentation networks. Since the quality of CAMs has a great influence on the segmentation results, recently, many strategies have been proposed to advance the original CAM method, including adversarial erasing [22, 53, 63, 66], online attention accumulation [25, 26], seed region expansion [24, 28], and affinity learning [1, 2, 57], *etc*. Despite the good performance, these works mostly take the whole input image as the sole input to the model. However, we empirically observe that classification models can discover more discriminative regions when taking local image patches as input compared to the whole input image. This suggests a proper way to improve the quality of attention maps by making use of local image patches.
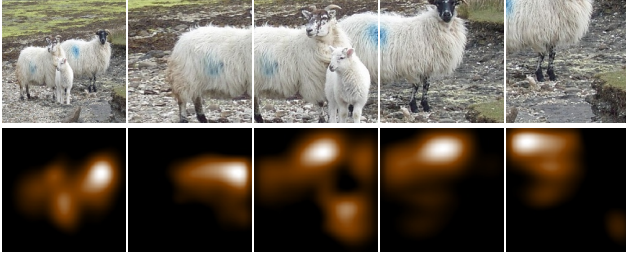
Figure 2. Motivation of our L2G attention knowledge transfer method. The top row shows the original image (global view) and multiple image patches after random crop (local views). The second row shows the attention maps generated by CAM [65]. We can observe that the attention maps of the local views capture more object details compared to that of the global view.

In this paper, taking the above analysis into account, we present a simple online local-to-global knowledge transfer framework, termed L2G, for generating high-quality object attentions. A conceptual illustration has been depicted in Fig. 1. Different from the aforementioned attention mining strategies, we propose to take advantage of both the global view and the local views randomly cropped from the input image (regions enclosed by the colorful bounding boxes). Specifically, our framework contains a local network that produces local attentions with rich object details for local views as well as a global network that receives the global view as input and aims to distill the discriminative attention knowledge from the local network.

Our method offers the following advantages. First of all, we produce attention maps from multiple local views of the input image rather than its global view. This allows us to attain more details on undiscovered semantic regions, which are also complementary across different local views, as shown in Fig. 2. Second, by designing a knowledge transfer loss, the complementary attention knowledge can be efficiently transferred to the global network in an online learning manner. This enables the global network to capture pixel-level semantic object details and produce high-quality attention maps in inference. Last but not the least, the overall pipeline is simple and flexible. We can selectively add additional constraints [32] to the local network to help shape the attained object attentions.

We evaluate our method on the PASCAL VOC 2012 and MS COCO 2014 datasets. Experiments demonstrate that our method achieves better performance than previous state-of-the-art methods. When using the DeepLab-v2 model [10] as our segmentation network, we attain 72.1% and 71.7% mIoU scores on the validation set and the test set of PASCAL VOC 2012, and 44.2% on the validation set of MS COCO 2014, setting new state-of-the-art records under the weakly supervised setting. We also conduct a series of ablation experiments to help readers better understand how each component performs in our method.

2. Related Work

2.1. Weakly Supervised Semantic Segmentation

One-stage WSSS methods directly utilize the image-level labels as supervision to train an end-to-end segmentation network. Early works [43, 44] formulate this problem as multiple instance learning. Later, Papandreou et al. [42] proposed an Expectation-Maximization (EM) method that utilizes the intermediate prediction to supervise the segmentation network. Zhang et al. [62] utilized the image classification branch to generate attention maps and constructed the pseudo segmentation labels to supervise the parallel segmentation branch. Araslanov et al. [3] proposed a self-supervised mechanism that applies the image appearance priors to generate pseudo segmentation labels during training. Chen et al. [7] constructed an end-to-end framework that uses an encoder-decoder network to explore object boundaries. Compared to two-stage WSSS methods, one-stage methods usually have inferior performance and are less attractive.

Two-stage WSSS methods rely on attention maps [63, 65] to generate pseudo segmentation labels, which are then used to train segmentation networks. The core of two-stage WSSS methods is to produce high-quality attention maps [5, 33, 48, 52, 54]. Towards this goal, a lot of works have been proposed recently. Wei et al. [53] proposed the adversarial erasing strategy, which iteratively occludes the mined object regions to drive the classification network to discover new object regions. Hou et al. [22] improved the adversarial erasing strategy by using a self-erasing strategy to prevent attention from spreading to the background. Kolesnikov et al. [28] introduced the seed-expansion idea, which expands the initial seed regions from the pre-computed attention maps and constrains the expanded regions to align with the object boundaries. Later, Jiang et al. [26] proposed the online attention accumulation strategy that utilizes the attention maps of different training phases. Chang et al. [6] exploited the sub-category information to highlight the non-discriminative semantic regions.

Another line of works attempts to refine attention maps to obtain integral object regions with precious boundaries. Ahn et al. [2] learned pixel affinity to propagate the semantics of strong responses in attention maps to the adjacent pixels. Chen et al. [8] and Ahn et al. [1] further improved this method by explicitly learning the class boundaries. Lee et al. [32] utilized the off-the-shelf saliency maps as supervision to guide the region learning to generate high-quality attention maps.

One common point shared by the aforementioned methods is that they all refine attention maps on the image's global view. Differently, our method takes advantage of both the global view and multiple local views and stud-
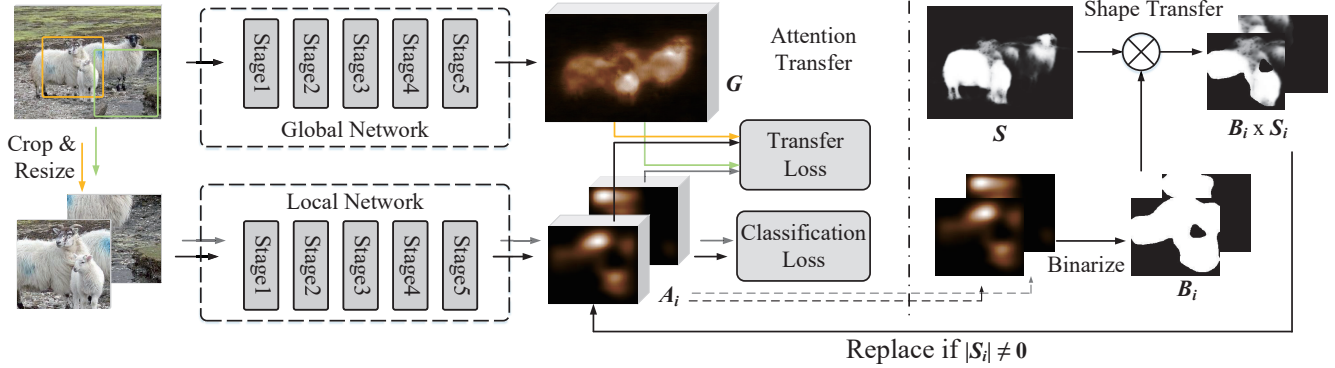
Figure 3. Overall framework of the proposed method. The complementary attention maps captured by the local network is distilled into the global network by a knowledge transfer loss.

ies how to efficiently transfer the complementary attention knowledge from the local network to the global network to improve the quality of attention maps.

2.2. Knowledge distillation

Our work is also related to knowledge distillation [17, 21, 61], which aims to distill the knowledge from the well-trained teacher model to a student model. For the image classification task, these works focus on improving the student model by imitating the prediction distribution of the teacher model. Moreover, some researchers [20, 39] also study knowledge distillation for the semantic segmentation task. Differently, we investigate how to transfer the attention knowledge captured by the local views to the global network in an online learning manner to better leverage the complementary information from multiple views.

3. Method

In this section, we present the whole framework of our method in detail. Before describing the framework, we first give some fundamental introduction to attention map generation.

3.1. Prerequisites

We first present the way to generate attention maps. Given an input image I, let y be the image-level label. The output feature F of the last convolutional layer has C channels, identical to the number of classes. The last convolutional layer is followed by a global average pooling layer, where the feature F is pooled to a vector f^C of size C. We calculate the classification loss by applying a sigmoid cross-entropy loss function, which is formulated as follows:

$$L_{\text{ce}} = -\frac{1}{C} \sum_{c=1}^{C} y^c \log(\sigma(f^c)) + (1 - y^c) \log(1 - \sigma(f^c)),$$

(1)

where σ is the sigmoid function. The attention maps can be generated from the output of the last convolutional layer. For some class c, the attention map A^c is derived from the cth channel of F, which can be formulated as

$$A^c = \frac{\text{ReLU}(F^c)}{\max(\text{ReLU}(F^c))}.$$

(2)

The above method, as pointed out in most previous work [2, 22, 26, 53], can only locate the most discriminative regions. It often fails in discovering those non-discriminative object regions that are semantically meaningful as well. In the following, we propose a novel attention generation framework by presenting a new local-to-global knowledge transfer method to capture high-quality object attentions.

3.2. Overall Framework

As mentioned in Sec. 1, the local network focusing on processing local patch views tends to discover more discriminative object regions. Based on this observation, we propose to leverage the attention maps for local views to aid a global network to locate more integral object regions.

The overall framework of the proposed approach can be found in Fig. 3. Functionally, there are four components: a global network, a local network, an attention transfer module, and a shape transfer module. The global network and the local network can be any CNN classifier, such as the popular VGGNet [46] or ResNet-38 [56]. In the attention transfer module, we optimize two loss functions: a classification loss L_{cls} that is used to recognize the semantic objects and an attention transfer loss L_{at} that encourages the global network to imitate the local network to discover more discriminative regions. In the shape transfer module, we introduce a shape constraint to loss L_{at}, yielding L_{st}, to shape the captured object attentions. Therefore, the overall optimized loss function can be formulated as follows:

$$L = L_{\text{cls}} + L_{\text{kt}}.$$

(3)

When no shape constraint is added, $L_{kt} = L_{at}$. Otherwise, $L_{kt} = L_{st}$.

3.3. Local-to-Global Attention Transfer

Given an input image I, we transform it into a set of different views V, including a global view V_I, and N local views $\{V_1, V_2, ..., V_N\}$, which are randomly cropped from the global view. The local views $\{V_1, V_2, ..., V_N\}$ are sent into the local network focusing on generating attention maps that contain rich object details. The global view V_I is fed into the global network, which aims to learn the knowledge from the local network and produces object attentions in inference. Let $\{F_1, F_2, ..., F_N\}$ be the outputs of the last convolutional layer of the local network and each has C channels corresponding to the number of classes. Let $\hat{F}$ be the output of the last convolutional layer of the global network that has $C+1$ channels. The classification loss and the attention transfer loss can be defined as follows.

Classification Loss: The classification loss is equipped with on the local network. Specifically, the feature maps $\{F_1, F_2, ..., F_N\}$ of the local views are first sent to a global pooling layer, where the features are pooled to a set of 1D feature vectors $\{f_1, f_2, ..., f_N\}$. Given a 1D feature vector f_i, the predicted probabilities for all categories can be computed by $q_i = \sigma(f_i)$. Recall that σ is the sigmoid function. Then, the classification loss L_{cls} can be written as

$$L_{cls} = -\frac{1}{N \times C} \sum_{i=1}^{N} \sum_{c=1}^{C} y^c \log(q_i^c) + (1 - y^c) \log(1 - q_i^c). \tag{4}$$

Attention Transfer Loss: We first generate attention maps for the local views from the local network. We use Eqn. (2) to generate attention maps $\{A_1^c, A_2^c, ..., A_N^c\}$ for the cth category if c is in the image-level labels. If c is not in the image-level labels, the attention values in the corresponding attention map will be zeroed. To transfer the attentions attained by the local network to the global network, we adopt the mean squared error loss.

Given the output $\hat{F}$ from the global network, we apply a Softmax function to $\hat{F}$ along the channel dimension for each location, yielding

$$G^c = \frac{e^{\hat{F}^c}}{\sum_{i=1}^{C+1} e^{\hat{F}^i}}, \tag{5}$$

where the value at each location of G^c means the probability of this location being category c. Let $\{G_1, G_2, ..., G_N\}$ denote the corresponding regions to $\{A_1, A_2, ..., A_N\}$ on the global view, *i.e.*, each pair (G_1, A_1) are cropped from the same coordinate on the global view. The attention transfer loss is formulated by measuring the difference between

$\{A_i\}$ and $\{G_i\}$ as follows:

$$L_{at} = \frac{1}{N} \sum_{i=1}^{N} ||A_i - G_i||^2. \tag{6}$$

During training, we jointly optimize the above two losses. During inference, the attention maps are generated from the global network while the local network can be discarded.

Discussion: Our method provides an efficient way to leverage the complementary information from the global view and the local views. The local-to-global attention transfer method conducts the global network to absorb the rich object detail knowledge captured by the local network in an online learning manner. Though most previous works also use data augmentations, like random crop, for the inputs, they do not have a component to accumulate the object detail knowledge from the cropped local patches online from a global view. This makes our local-to-global strategy quite different from previous works. We will show more advantages of the proposed approach over other methods in the experiment section.

3.4. Local-to-Global Shape Transfer

The proposed local-to-global attention transfer strategy can already result in more integral object attentions than the original CAM [65]. However, as the attention transfer process leverages only the image-level labels, the captured attentions around the object boundaries are not sharp enough. To well capture the shape of the localized objects in the attention maps, we attempt to introduce auxiliary salient object information into the attention transfer loss by adding a shape constraint. The saliency model [38] can serve as a class-agnostic salient object detector, which can segment the foreground objects and provide shape information.

The shape transfer process is simple, which has been illustrated in the right part of Fig. 3. Given the attention maps $\{A_i\}$ from the local network, we first binarize them with a small threshold (e.g., 0.1), yielding the binary maps $\{B_i\}$. Then, we utilize the saliency model to generate the saliency map S for the given image I and get the corresponding saliency regions to the attention maps $\{A_i\}$ with the same coordinate on I denoted as $\{S_i\}$. The attention transfer loss can then be rewritten as

$$L_{st} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} ||B_i \times S_i - G_i||^2, & \text{if } |S_i| \neq 0 \\ ||A_i - G_i||^2, & \text{if } |S_i| = 0 \end{cases} \tag{7}$$

where $\times$ denotes element-wise multiplication and $|S_i|$ is the cardinality of the saliency map S_i. By using $B_i \times S_i$, we aim to remove the attention regions outside the salient objects, which belong to the background with high probability. This allows our method to fully leverage the shape information provided by the saliency maps and results in high-quality
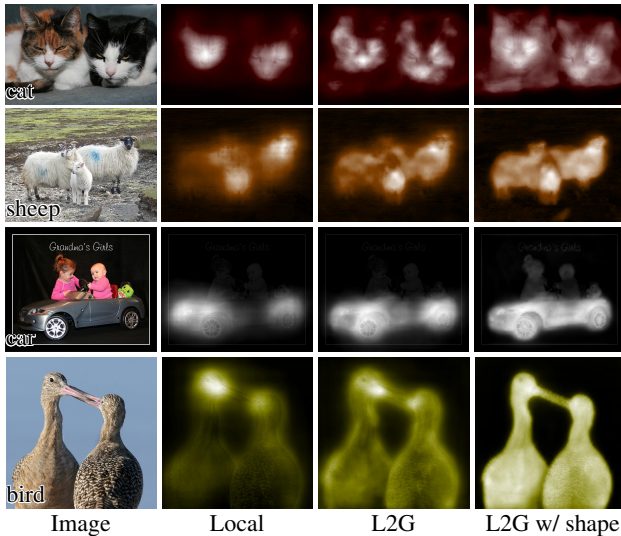
Figure 4. Qualitative comparison of attention maps from different methods.

Figure 5. Ablations on the local view size and number N.

attention maps. We will elaborate more on this in our experiment section. Note that as not all the images would have salient objects, it is inappropriate to always use the top part of Eqn. (7). Thus, for those images whose saliency maps contain nothing, we utilize the original attention maps as supervision as formulated in the bottom part of Eqn. (7).

It is worth mentioning that EPS [32] also uses saliency maps as supervision to provide the network with shape information. Differently, our method focuses more on how to take advantage of multiple local views and how to efficiently transfer the learned knowledge from the local network to the global one. In the following, we will show the advantage of the proposed local-to-global knowledge transfer over EPS.

4. Experiments

4.1. Experimental Setup

Datasets and evaluation metric. Experiments are conducted on two publicly available datasets, PASCAL VOC 2012 and MS COCO 2014. The PASCAL VOC 2012 dataset contains 20 semantic categories and the background. It is split into three sets, the training, validation, and test sets, each containing 1464, 1449, and 1456 images, respectively. Following most previous works, we also use the augmented training set [18], yielding totally 10582 training images. The MS COCO 2014 dataset has 80 semantic categories. Following [11, 32], the images without target categories are excluded from the dataset, remaining 82081 training images and 40137 validation images. The mean intersection-over-union (mIoU) [41] is used as the evaluation metric.
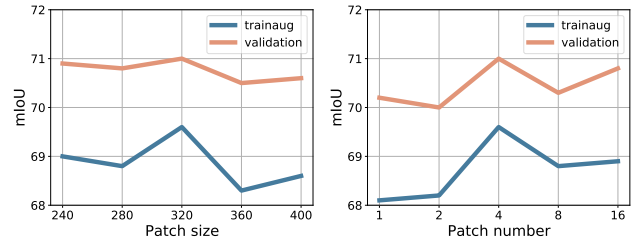
Network settings. For data augmentation, the short size of the input image is resized to 512. The global view is with a resolution of $448{\times}448$, which is cropped from the input image. The local image patches with resolution $320{\times}320$ are cropped from the global view. Following [2, 32], we utilize ResNet-38 [56] as our classification network. We train the classification network for 10 epochs on PASCAL VOC and 15 epochs on MS COCO, respectively. The initial learning rate is set to 0.001. The attention maps are generated from the global network using the multi-scale test strategy [2]. We select DeepLab-v1 [9] and DeepLab-v2 [10] as our segmentation networks. We report performance based on both VGG-16 [46] and ResNet-101 [19]. For VGG-16 based segmentation network, we use the classification model pretrained on ImageNet [13] for initialization. For ResNet-101, we use the COCO pretrained model. For the experiments on MS COCO dataset, we all utilize ImageNet pretrained model. Following [32], we use the same way to generate the pseudo labels. Given the attention maps, we assign a fixed threshold to the background channel and use the argmax function to obtain the label for each pixel.

4.2. Ablation Study

Local view sampling strategy. First, we study the impact of the sampling strategy on the attention maps. We compare two local view sampling strategies. One is the random sampling strategy, and the other is the uniform sampling strategy. We implement the uniform sampling strategy by sliding the window over the global view uniformly. In this way, every pixel can be enclosed within some local view. For the global view with $448{\times}448$ resolution, we set the window size to $320{\times}320$ and the stride to 64, obtaining 9 local views. For a fair comparison with the uniform sampling strategy, we randomly sample 9 image patches for the random sampling strategy. The qualities of the pseudo segmentation labels using these two strategies are quite close to each other (random 68.8% v.s. uniform 68.5%). To flexibly adjust the local view number N, we choose the random sampling strategy in our method.

Patch size and patch number N. The patch size controls the spatial size of the local views. The patch number N denotes the number of the local views sent to the local net-

Table 1. Comparisons of mIoU scores under different network settings. The baseline is the original CAM [65]. SW: The sliding window strategy is applied to the baseline during inference [65]. Local: Using multiple local image patches instead of the input image to train the classification network. L2G: Our method with local-to-global attention transfer only. $mIoU_{trainaug}$ denotes the mIoU score of the pseudo segmentation labels on the augmented training set.

No.	SW	Local	L2G	$mIoU_{trainaug}$	$mIoU_{val}$
1				47.1	47.5
2	✓			46.1 (**-1.0**)	47.2 (**-0.3**)
3		✓		48.5 (**+1.4**)	50.0 (**+2.5**)
4			✓	56.8 (**+9.7**)	54.9 (**+7.4**)

Table 2. Ablation experiments on the importance of each component. L2G: Local-to-global attention transfer only. Shape: Local-to-global shape transfer. As can be seen, our local-to-global transfer strategy can significantly improve the performance compared to the setting with only the local network being used. When the shape information is incorporated, L2G can still lift the performance by a large margin.

No.	Local	L2G	Shape	$mIoU_{trainaug}$	$mIoU_{val}$
1	✓			48.5	50.0
2		✓		56.8 (**+8.3**)	54.9 (**+4.9**)
3	✓		✓	68.0	69.9
4		✓	✓	70.3 (**+2.3**)	72.1 (**+2.2**)

work. To study their impact on the attention quality, we select 5 different patch sizes [240×240, 280×280, 320×320, 360×360, 400×400]. When studying the patch number N, we select the number of local views from the range of [1, 2, 4, 8, 16]. As shown in Fig. 5, we observe that when N increases, the quality of the pseudo segmentation labels becomes better. The performance tends to be robust when the local view number is larger than 4. For the patch size, we can see that our method achieves the best performance when the local view size takes 320×320. When the size is larger than 320×320, the quality of the pseudo segmentation labels decreases largely.

Importance of the proposed local-to-global knowledge transfer. When sending the local views to the local network, we can discover more object regions from the resulting attention maps. Here, one may raise a question: "Are the attention maps from the local network good enough so that we do not need the transfer process?". To answer this question, we test the quality of the pseudo segmentation labels using the attention maps from the local network. As shown in Tab. 1, we can see the performance of the local network is slightly better than the baseline CAM [65]. However, the performance is far lower than L2G (48.5% v.s. 56.8%). We also show some qualitative results of the attention maps in Fig. 4 and segmentation results in Fig. 6. This indicates the local-to-global attention transfer strategy is a more efficient way to leverage the rich object attention knowledge captured by the local network.

In addition, we further extend the above experiments by introducing the shape information. The corresponding results can be found in Tab. 2. It has been demonstrated in [32] that saliency shape information can significantly improve attention quality. However, when the proposed L2G is used, the mIoU scores on both the trainaug and validation sets can be largely improved. We will show more numerical results on segmentation in the next subsection.

L2G v.s. Sliding window. The key of our method is to leverage the local attention maps to facilitate the global net-

work to discover more integral object regions. One direct way to implement this idea is to utilize the sliding window strategy during inference and aggregate the attention maps from different image patches. We compare our L2G with the sliding window strategy. Specifically, for the sliding window strategy, the window size and the stride are set to 320×320 and 64, respectively. For our L2G, we set the local view size to 320×320 and the number of randomly sampled patches to 4 at a time.

As shown in Tab. 1, the local-to-global attention transfer strategy achieves much better results than the baseline CAM [65], which verifies the effectiveness of our method. However, the results of the sliding window strategy are even worse than the original CAM. We argue that the sliding window strategy is not suitable for mining non-discriminative object regions as the trained model is still based on inputs with the global view. This makes the undiscovered object regions that have different appearances with the distinctive regions hard to respond when processing the global view.

Classification loss in the global network. The local network is equipped with the classification loss to guide the attention generation. One may ask the question "Does the global network also need the classification loss?" To answer this question, we have attempted to add a classification loss to the global network. We observe that the attention maps generated from the global network locate very small object regions when adding the classification loss. The quality of the pseudo segmentation labels decreases largely from 70.3% to 53.8%. We argue that the classification loss and the attention transfer loss play opposite roles. The classification loss makes the attention be more discriminative. The attention transfer loss aids in transferring the attention on non-discriminative regions to the global network. Thus, the attention maps become worse.

Local and global backbone sharing. Here, we explore the performance gap between with/without the local and global network backbone sharing. When the local and global networks share the same backbone, the mIoU score of the
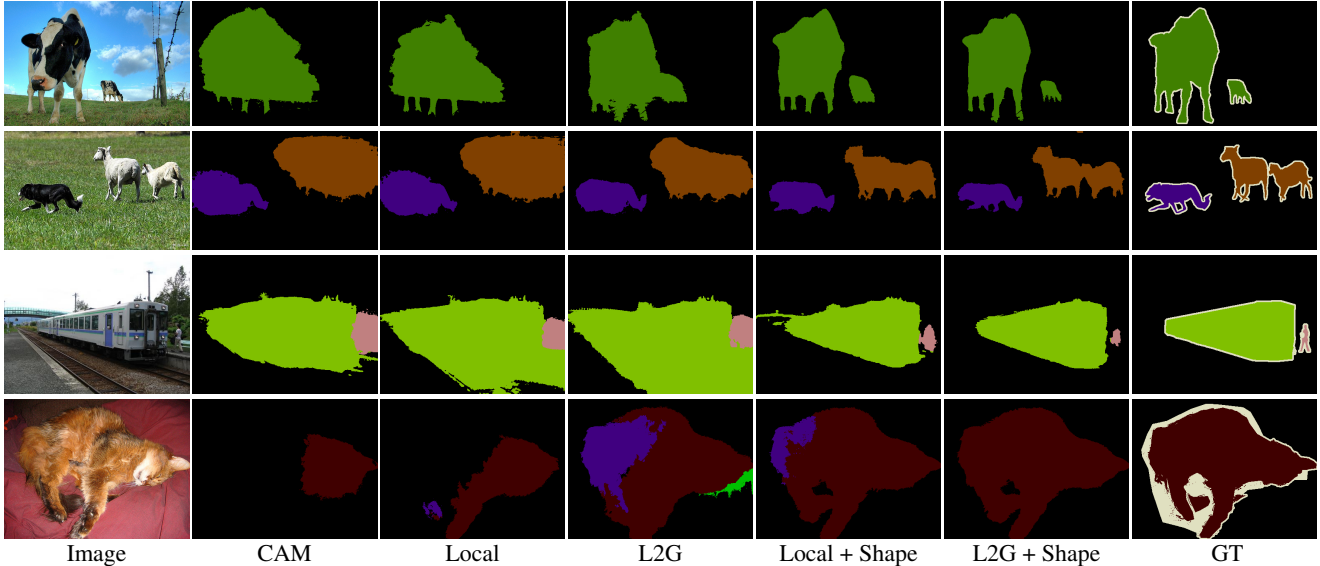
Figure 6. Comparison of the segmentation results under different network settings. We can observe that the combining L2G and shape transfer yields the best results, especially on local object details.

| Image | CAM | Local | L2G | Local + Shape | L2G + Shape | GT |

Table 3. Comparison of pseudo segmentation labels on the PASCAL VOC train set with no saliency maps.

Methods	$mIoU_{train}$
CAM [65]	48.0
SC-CAM [6]	50.9
SEAM [52]	55.4
ADvCAM [31]	55.6
L2G (ours)	56.2

Table 4. Comparisons of pseudo segmentation labels on the PASCAL VOC train set with saliency maps incorporated.

Methods	$mIoU_{train}$
SGAN [58]	62.8
EPS [32]	69.4
L2G (ours)	71.9

Table 5. Quantitative comparisons to previous state-of-the-art approaches on PASCAL VOC 2012 validation and test sets. All the segmentation results are based on the DeepLab with VGGNet backbone [46]. Pub.: Publication, Seg.:Segmentation network, Sup.: Supervision, I.: Image-level label, S.: saliency maps from the off-the-shelf saliency model.

Methods	Pub.	Seg.	Sup.	Val (%)	Test (%)
AffinityNet [2]	CVPR'18	V1	I.	58.4	60.5
MCOF [51]	CVPR'18	V1	I.+S.	56.2	57.6
DSRG [24]	CVPR'18	V2	I.+S.	59.0	60.4
SeeNet [22]	NeurIPS'18	V1	I.+S.	61.1	60.7
FickleNet [30]	CVPR'19	V2	I.+S.	61.2	61.9
OAA$^+$ [26]	ICCV'19	V1	I.+S.	63.1	62.8
BES [8]	ECCV'20	V1	I.	60.1	61.1
MCIS [48]	ECCV'20	V1	I.+S.	63.5	63.6
Multi-Est. [16]	ECCV'20	V1	I.+S.	64.6	64.2
ICD [15]	CVPR'20	V1	I.+S.	64.0	63.9
ECS-Net [49]	ICCV'21	V1	I.	62.1	63.4
DRS [27]	AAAI'21	V1	I.+S.	63.5	64.5
Group-WSSS [34]	AAAI'21	V2	I.+S.	63.3	63.6
OAA++$^+$ [25]	PAMI'21	V1	I.+S.	63.7	63.2
NSROM [59]	CVPR'21	V2	I.+S.	65.5	65.3
EPS [32]	CVPR'21	V1	I.+S.	66.6	67.9
EPS [32]	CVPR'21	V2	I.+S.	67.0	67.3
L2G (ours)	–	V1	I.+S.	68.1	68.8
L2G (ours)	–	V2	I.+S.	68.5	68.9

pseudo segmentation labels on the trainaug set is 69.2%. After training the segmentation network, the mIoU score on the validation set is 70.9%. When the local and global networks utilize different backbones, the mIoU score of the pseudo segmentation labels can be improved by 1.1%. The final segmentation result also attains 1.2% mIoU gains.

4.3. Comparisons with the State-of-the-Arts

We first compare the quality of our produced attention maps with the previous state-of-the-art WSSS methods. Our attention maps are converted to pseudo segmentation labels. As shown in Tab. 3 and Tab. 4, it is obvious that the attention maps generated by our method are better than other methods no matter whether the saliency maps are used. Without the saliency maps, the mIoU score on the PASCAL VOC train set reaches 56.2%, better than SEAM [52] by 0.8%. After applying the saliency maps to the transfer process, the mIoU score reaches 71.9%, much better than EPS [32] (71.9% v.s. 69.4%).

We use the pseudo segmentation labels to train the

DeepLab segmentation model directly. We compare the segmentation performance of our method with previous state-of-the-art methods. Tab. 5 and Tab. 6 list the segmentation results of our method and the recent state-of-the-art

Table 6. Quantitative comparisons to previous state-of-the-art approaches on PASCAL VOC 2012 validation and test sets. All the segmentation results are based on the ResNet backbone [19, 56]. Our method achieves the best results.

Methods	Publication	Seg.	Sup.	Val (%)	Test (%)
AffinityNet [2]	CVPR'18	V1	I.	61.7	63.7
MCOF [51]	CVPR'18	V1	I.+S.	60.3	61.2
DSRG [24]	CVPR'18	V2	I.+S.	61.4	63.2
SeeNet [22]	NeurIPS'18	V1	I.+S.	63.1	62.8
IRNet [1]	CVPR'19	V1	I.	63.5	64.8
FickleNet [30]	CVPR'19	V2	I.+S.	64.9	65.3
OAA$^+$ [26]	ICCV'19	V1	I.+S.	65.2	66.4
SSDD [45]	ICCV'19	V1	I.	66.1	66.8
SEAM [52]	CVPR'20	V2	I.	64.5	65.7
SC-CAM [6]	CVPR'20	V2	I.	66.1	65.9
ICD [15]	CVPR'20	V1	I.+S.	67.8	68.0
BES [8]	ECCV'20	V2	I.	65.7	66.6
MCIS [48]	ECCV'20	V1	I.+S.	66.2	66.9
Multi-Est. [16]	ECCV'20	V1	I.+S.	67.2	66.7
LIID [40]	PAMI'20	V2	I.+IS.	66.5	67.5
DRS [27]	AAAI'21	V2	I.+S.	71.2	71.4
Group-WSSS [34]	AAAI'21	V2	I.+S.	68.2	68.5
ECS-Net [49]	ICCV'21	V1	I.	66.6	67.6
PMM [35]	ICCV'21	PSP	I.	68.5	69.0
CDA [47]	ICCV'21	V2	I.	66.1	66.8
CGNet [29]	ICCV'21	V1	I.	68.4	68.2
AuxSegNet [57]	ICCV'21	V1	I.+S.	69.0	68.6
AdvCAM [31]	CVPR'21	V2	I.	68.1	68.0
NSROM [59]	CVPR'21	V2	I.+S.	70.4	70.2
EDAM [55]	CVPR'21	V1	I.+S.	70.9	70.6
EPS [32]	CVPR'21	V1	I.+S.	71.0	71.8
EPS [32]	CVPR'21	V2	I.+S.	70.9	70.8
L2G (ours)	–	V1	I.+S.	72.0	73.0
L2G (ours)	–	V2	I.+S.	72.1	71.7

Table 7. Quantitative comparisons to previous state-of-the-art approaches on MS COCO validation set. All the segmentation results are based on VGGNet backbone [46] except L2G* using ResNet-101 backbone [19].

Methods	Publication	Seg.	Sup.	Val (%)
SEC [28]	ECCV'16	V1	I.+S.	22.4
DSRG [24]	CVPR'18	V2	I.+S.	26.0
ADL [11]	PAMI'20	V1	I.+S.	30.8
EPS [32]	CVPR'21	V2	I.+S.	35.7
L2G (ours)	–	V2	I.+S.	42.7
L2G* (ours)	–	V2	I.+S.	44.2

improve the results. Thus, we believe there is still a large room to improve the proposed framework, and we also hope our local-to-global knowledge transfer method could provide researchers with a new research direction.

Analysis of failure cases. First, some non-target objects are wrongly recognized as the target classes as shown in the first row of Fig. 7. In our L2G, we only use ResNet to extract attention. Designing more advanced classification models, such as transformers [14, 60], could improve the results. Second, the shape of the discovered objects is still being further improved (the second row). Using stronger saliency models or over-segmentation methods could, to some extent, solve this.
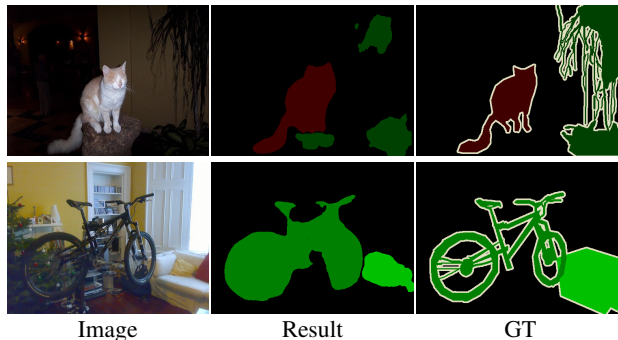

Figure 7. Two failure segmentation examples of our L2G.

methods on the PASCAL VOC dataset. As we can see, compared to the previous WSSS methods, our method achieves the best results on both the validation and test sets. The work most relevant to our method is EPS [32], which explicitly uses the saliency maps as supervision. The differences between our method and EPS have been explained in Sec. 3.4. As shown in Tab. 6, we can see that our method can improve the results of EPS by around 1%. Besides, as shown in Tab. 7, our results on the challenging MS COCO dataset are much better than the previous methods, which also demonstrates the effectiveness of our local-to-global strategy. The mIoU of the pseudo labels for our method is 43.4%, much better than that of EPS (37.2%).

Discussion. It is worthy to note that our local network is just a simple classification model. Because of the flexibility of the proposed framework, we can replace the local network with more complicated attention models to further

5. Conclusion

In this paper, we propose a novel local-to-global attention transfer method to attain object attentions. By leveraging the complementary attention captured by the local network from the local views and introducing the shape constraint to the attention transfer process, our method achieves the best results on both the validation and test sets of PASCAL VOC 2012 and the validation set of MS COCO 2014. We hope the proposed approach could facilitate the research on vision tasks relying on high-quality attention maps.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2209–2218, 2019.

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4981–4990, 2018.

[3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4253–4262, 2020.

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Eur. Conf. Comput. Vis.*, pages 549–565, 2016.

[5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Mixupcam: Weakly-supervised semantic segmentation via uncertainty regularization. In *Brit. Mach. Vis. Conf.*, 2020.

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8991–9000, 2020.

[7] Jianjun Chen, Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Yue Hu, and Jianlong Tan. End-to-end boundary exploration for weakly-supervised semantic segmentation. In *ACM Int. Conf. Multimedia*, pages 2381–2390, 2021.

[8] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Eur. Conf. Comput. Vis.*, pages 347–362. Springer, 2020.

[9] 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 *Int. Conf. Learn. Represent.*, 2015.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.

[11] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[12] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 1635–1643, 2015.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.

[15] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4283–4292, 2020.

[16] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 332–348. Springer, 2020.

[17] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Int. Conf. Mach. Learn.*, pages 1607–1616, 2018.

[18] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Int. Conf. Comput. Vis.*, pages 991–998, 2011.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[20] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 578–587, 2019.

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2014.

[22] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, volume 31, pages 549–559, 2018.

[23] Qibin Hou, Daniela Massiceti, Puneet Kumar Dokania, Yunchao Wei, Ming-Ming Cheng, and Philip HS Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In *Int. Worksh. on Energy Minimization Methods in Comput. Vis. Pattern Recog.*, pages 263–277, 2017.

[24] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7014–7023, 2018.

[25] Peng-Tao Jiang, Ling-Hao Han, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Online attention accumulation for weakly supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[26] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Int. Conf. Comput. Vis.*, pages 2070–2079, 2019.

[27] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *AAAI Conf. Artif. Intell.*, pages 1754–1761, 2021.

[28] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Eur. Conf. Comput. Vis.*, pages 695–711, 2016.

[29] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 6994–7003, 2021.

[30] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5267–5276, 2019.

[31] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4071–4080, 2021.

[32] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5495–5505, 2021.

[33] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9215–9223, 2018.

[34] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *AAAI Conf. Artif. Intell.*, pages 1984–1992, 2021.

[35] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 6964–6973, 2021.

[36] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3159–3167, 2016.

[37] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1925–1934, 2017.

[38] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3917–3926, 2019.

[39] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2604–2613, 2019.

[40] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015.

[42] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Int. Conf. Comput. Vis.*, pages 1742–1750, 2015.

[43] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *Int. Conf. Learn. Represent.*, 2015.

[44] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1713–1721, 2015.

[45] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 5208–5217, 2019.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2015.

[47] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, 2021.

[48] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 347–365, 2020.

[49] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Int. Conf. Comput. Vis.*, pages 7283–7292, 2021.

[50] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7158–7166, 2017.

[51] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1354–1362, 2018.

[52] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12275–12284, 2020.

[53] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1568–1576, 2017.

[54] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7268–7277, 2018.

[55] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16765–16774, 2021.

[56] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recogn.*, 90:119–133, 2019.

[57] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary

tasks with affinity learning for weakly supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 6984–6993, 2021.

[58] Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020.

[59] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2623–2632, 2021.

[60] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Int. Conf. Comput. Vis.*, pages 558–567, 2021.

[61] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3903–3911, 2020.

[62] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *AAAI Conf. Artif. Intell.*, pages 12765–12772, 2020.

[63] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1325–1334, 2018.

[64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2881–2890, 2017.

[65] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2929, 2016.

[66] Kuangqi Zhou, Qibin Hou, Zun Li, and Jiashi Feng. Multi-miner: Object-adaptive region mining for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2006.07834*, 2020.