Integral Object Mining via Online Attention Accumulation

Peng-Tao Jiang[1] Qibin Hou[1] Yang Cao[1] Ming-Ming Cheng[1*]
Yunchao Wei[2] Hong-Kai Xiong[3]
[1]TKLNDST, CS, Nankai University [2]UTS [3]Shanghai Jiaotong University
pt.jiang@mail.nankai.edu.cn cmm@nankai.edu.cn

Abstract

Object attention maps generated by image classifiers are usually used as priors for weakly-supervised segmentation approaches. However, normal image classifiers produce attention only at the most discriminative object parts, which limits the performance of weakly-supervised segmentation task. Therefore, how to effectively identify entire object regions in a weakly-supervised manner has always been a challenging and meaningful problem. We observe that the attention maps produced by a classification network continuously focus on different object parts during training. In order to accumulate the discovered different object parts, we propose an online attention accumulation (OAA) strategy which maintains a cumulative attention map for each target category in each training image so that the integral object regions can be gradually promoted as the training goes. These cumulative attention maps, in turn, serve as the pixel-level supervision, which can further assist the network in discovering more integral object regions. Our method (OAA) can be plugged into any classification network and progressively accumulate the discriminative regions into integral objects as the training process goes. Despite its simplicity, when applying the resulting attention maps to the weakly-supervised semantic segmentation task, our approach improves the existing state-of-the-art methods on the PASCAL VOC 2012 segmentation benchmark, achieving a mIoU score of 66.4% *on the test set. Code is available at* https://mmcheng.net/oaa/.

1. Introduction

Benefiting from the large-scale pixel-level training data and advanced convolutional neural network (CNN) architectures, fully-supervised semantic segmentation approaches, such as [4, 20, 22, 42, 38], have made great progress recently. However, constructing a large-scale pixel-accurate dataset is fairly expensive and requires considerable human efforts and time cost. In order to economize human labors, researchers propose to learn semantic segmentation using weak supervision, such as bounding boxes [27], points [2], and even image-level annotations [26]. Among these weak supervisions, image-level annotations can be more easily obtained than other annotations. Thus, in this paper, we focus on semantic segmentation under image-level supervision.

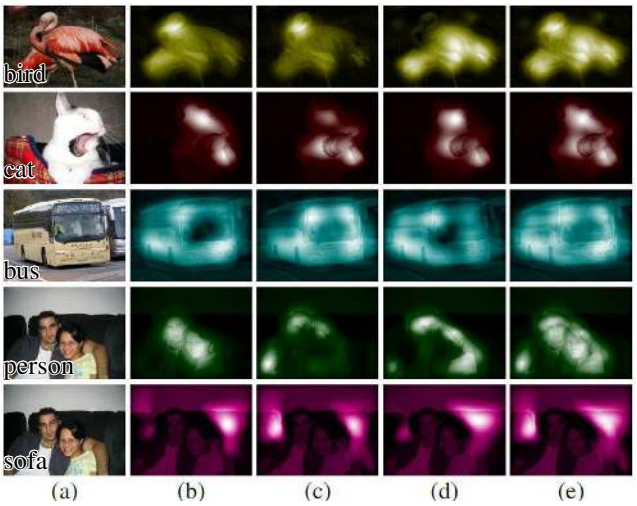

Figure 1. Observation of our proposed approach. (a) Source images; (b-d) Intermediate attention maps produced by a classification network at different training stages; (e) Cumulative attention maps produced by combining attention maps in (b), (c), and (d) through a simple element-wise maximum operation. It can be easily observed that the discriminative regions continuously shift over different parts of the semantic objects. The fused attention maps in (e) can record most of semantic regions compared to (b), (c), and (d). Best viewed in color.

Because of the ability to discover discriminative attention regions, classification models [43, 29] have been widely used in the weakly-supervised semantic segmentation task for generating initial class-specific seeds. However, the discovered regions often focus on small parts of the semantic objects, which limits the capability of segmentation networks to learn rich pixel-level semantic knowledge. Later methods consider leveraging the adversarial erasing strategy [35, 40]

*M.M. Cheng is the corresponding author.

to mine more semantic regions. Unfortunately, as the training process continues, the discriminative regions expand, and thus some undesired background stuff is also predicted as foreground. In [37], dilated convolution is revisited for attention generation. However, the convolution layers with larger dilation rates often lead to the appearance of noisy regions.

One common point shared by the above approaches is that they all utilize the final classification models to generate attention maps. In this paper, we consider the attention generation process from a new perspective. We observe that the discriminative regions discovered at different training stages constantly shift over different parts of the semantic objects before the classification network reaches convergence. The main reasons can be briefly summarized as follows:

- First, a powerful classification network usually seeks robust common patterns for a specific category so that all the images from such a category can be well recognized. Therefore, those training samples that are hard to be correctly classified will drive the network to make changes in choosing common patterns, leading to the continuous shift of attention regions until the network reaches convergence.

- Second, in the training phase, attention maps produced by the current attention model are mostly influenced by the previous input images. Therefore, images with different content and the input order of the training images will both lead to the variation of the discriminative regions in the intermediate attention maps.

More interestingly, we also observe that the discriminative regions discovered at different training phases are often complementary, which reflects the importance of leveraging the intermediate attention maps for detecting integral objects. Fig. 1(b-d) gives a clear illustration of this phenomenon, which shows the variation of attention regions as the training process continues. If these discriminative regions in the intermediate attention maps can be recorded, we may successfully promote the capability of detecting complete semantic objects with only image-level supervision.

Based on the above observation, we introduce a simple yet effective approach for attention generation, which is capable of taking the intermediate states of classification networks into account. Specifically, we present an online attention accumulation (OAA) strategy, in which a cumulative attention map for each category in each image is maintained to sequentially accumulate the discriminative regions produced by the classification network at different training phases. The complementarity of the intermediate attention maps enables discovering integral semantic objects to be possible (see Fig. 1e). Despite the relatively complete attention regions by OAA compared to CAM [43], some attention valuses in object regions are still not strong enough. To improve this situation, we further design a hybrid loss function (the combination of an enhanced loss and a constraint loss) to train an integral attention model by taking the cumulative attention maps as soft labels. In this way, the new attention model advances the OAA strategy and can generate more integral object regions. To evaluate the quality of the attention maps by our approach, we conduct a series of ablation experiments and apply them to the weakly-supervised semantic segmentation task. We show significant improvements over existing methods on the popular PASCAL VOC 2012 segmentation benchmark [8] (a mean IoU score of 66.4% on the test set). We hope the thought of OAA could promote the development of attention models or even other research areas in the future.

2. Related Work

In this section, we briefly review the history of attention models and describe the weakly-supervised semantic segmentation methods that are strongly related to our work.

2.1. Visual Attention

To date, some outstanding work has been proposed in order to get high-quality attentions. As an early attempt, Simonyan *et al*. [31] used the error back-propagation strategy to visualize semantic regions. Later, CAM [43] shows the ability of the global average pooling (GAP) layer by using it to convolutional neural networks to detect the class activation maps. Based on CAM, Grad-CAM [29] proposes a technique for producing visual explanations for any target concept such as image classification, VQA, and image captioning by flowing the gradients into the final convolutional layer to produce coarse attention maps. Moreover, some researchers were inspired by the top-down human visual attention system and proposed a new method called Excitation Backprop [39], which hierarchically propagated the top-down signals downwards in the network via a probabilistic Winner-Take-All process. Recently, different from the above methods for explaining the networks, some work [40, 19, 37, 14, 41] produced attention maps by localizing large and integral relevant regions of the semantic objects for weakly-supervised semantic segmentation. All the above methods utilize the final classification models to generate attention. Besides top-down visual attention, recent researches [34, 13, 37] also found that bottom-up salient objects cues [12, 33, 6] are very useful for extracting background cues.

2.2. Weakly-Supervised Semantic Segmentation

Weakly-supervised semantic segmentation has also experienced great progress as a variety of methods were proposed. Among these methods, we only introduce some segmentation approaches with image-level supervision that are strongly related to our work. The mainstream methods [18, 34, 15, 37, 1] use the attention maps as initial seeds.

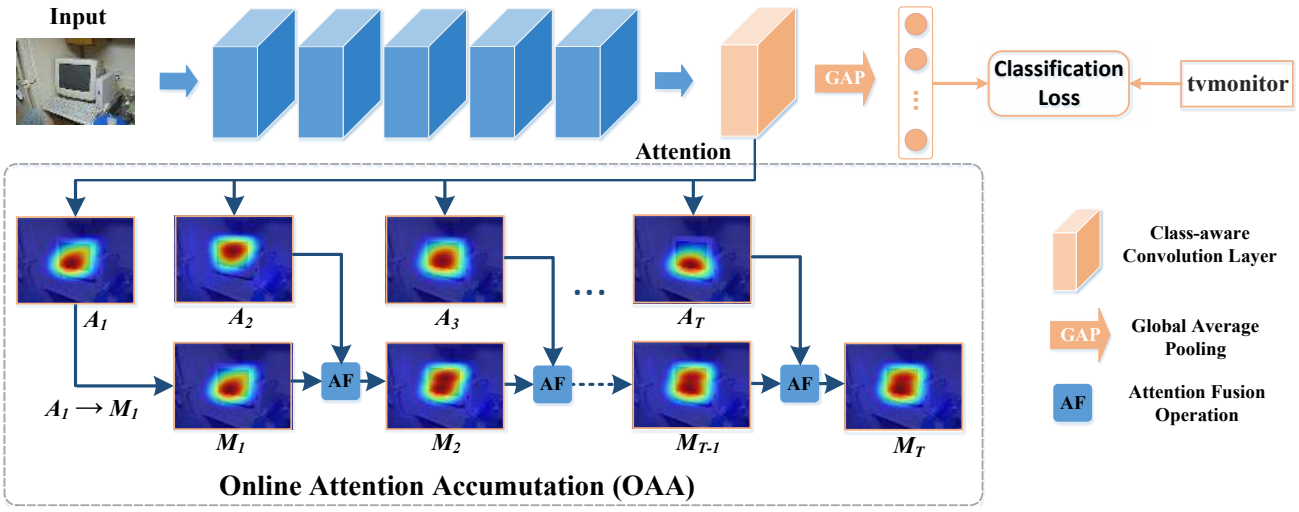

Figure 2. Illustration of our online attention accumulation (OAA) process. The attention maps are generated online from the class-aware convolutional layer. Our OAA utilizes these discriminative regions of attention maps at the different training phases and integrates them into the cumulative attention maps with a simple attention fusion strategy progressively.

Typically, SEC [18] introduced three loss functions called seeding, expansion and boundary constrain losses to expand the initial seeds and meanwhile train the segmentation model. However, the performance of these methods is limited in that the object-related seeds only cover small and sparse semantic regions.

More recently, researchers proposed a variety of methods to mine integral object regions based on classification networks. In [35], Wei *et al*. proposed an approach which uses an adversarial erasing (AE-PSL) strategy to mine different regions of the objects progressively in order to obtain dense maps. However, the procedures of AE-PSL are complicated, which requires repetitive training procedures and learns multiple classification models to obtain different object regions. GAIN [19] improved the adversarial erasing strategy by using attention maps to provide a self-guidance that forces the network to focus attention on the objects holistically.

3. Methodology

In this section, we describe the pipeline of our proposed approach and exhaustively explain the working mechanism of each component in our framework. Fig. 3 illustrates the whole framework of our method.

3.1. Attention Generation

In this paper, we adopt CAM [43] as our default discriminative region generator. In order to obtain attention maps at the training stage, we use the class-specific feature maps outputted by the last convolutional layer to generate attention maps, which is proven by [40] identical to the attention generation process in CAM.

The basic architecture can be found on the top of Fig. 2. Like most previous work [40, 37], we also adopt the VGG-16 [32] as our backbone. First, three convolutional layers are added on the top of the fully-convolutional backbone, each of which is followed by a ReLU layer for nonlinear transformation. A class-aware convolutional layer of C channels with kernel size 1×1 is then added for capturing the attention. Here C is the number of categories. Let F be the output of the class-aware convolutional layer. Regarding the fact that some images may have more than one category, we treat the whole training process as C binary classification problems. The probability of predicting the target category c can be computed by

$$p^c = \sigma\big(\mathrm{GAP}(F^c)\big), \tag{1}$$

where GAP is the global average pooling operation, and $\sigma(\cdot)$ is the sigmoid function. The cross-entropy loss is used to optimize the whole network. To get the attention maps given an image I, the feature map F is first fed into a ReLU layer, and then a simple normalization operation is performed to make sure the values in each attention map range from 0 to 1:

$$A^c = \frac{\mathrm{ReLU}(F^c)}{\max(F^c)}. \tag{2}$$

We then apply the attention maps generated at different training stages into the OAA process.

3.2. Online Attention Accumulation

To effectively implement our observation, we propose an online attention accumulation (OAA) strategy. When the training images are fed into the network at different training epochs, OAA combines the generated attention maps from the classification models. In particular, as shown in Fig. 2, for each target class c in a given training image I, we establish a cumulative attention map M^c which is used to preserve

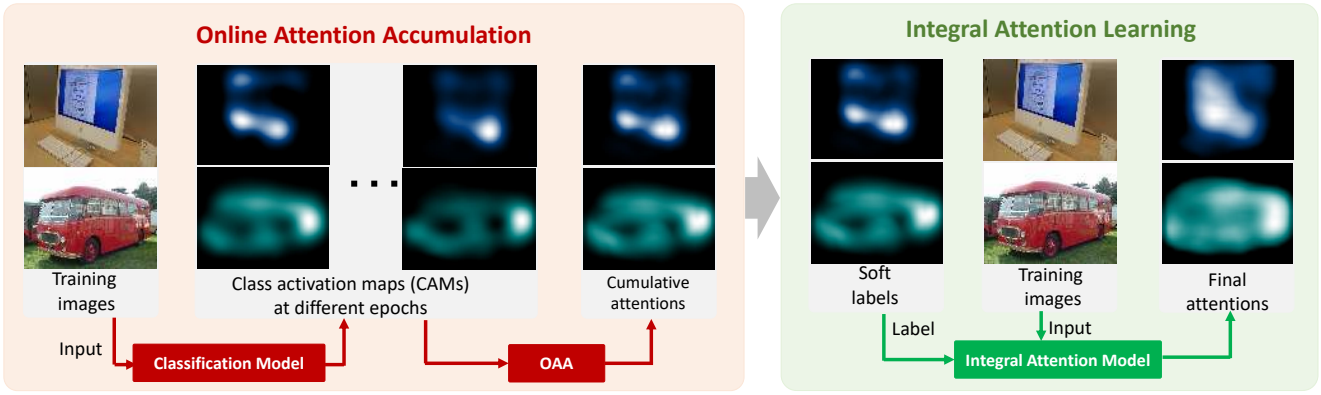

Figure 3. Pipeline of our OAA$^+$ approach. The attention maps generated by the classification network during different training time are fused into the cumulative attention maps to mine the object regions as entire as possible. Then the obtained cumulative attention maps are utilized as pixel-level supervision to train the integral attention model, which further advances the quality of the attention maps.

the discovered discriminative regions. Our OAA first uses the attention map A_1[1] of class c at the first epoch (i.e., A_1 is obtained when the training image is inputted to network for the first time) to initialize the cumulative attention map M_1. Then, when the image is inputted to the network for the second time, the OAA updates the cumulative attention map by combining M_1 and the newly generated attention map A_2 according to the following fusion strategy:

$$M_2 = \mathrm{AF}\big(M_1, A_2\big), \tag{3}$$

where $\mathrm{AF}(\cdot)$ represents the attention fusion strategy. Similarly, at the t-th epoch, the OAA uses the attention map A_t to update the cumulative attention map M_{t-1}, yielding

$$M_t = \mathrm{AF}\big(M_{t-1}, A_t\big). \tag{4}$$

The OAA repeats the above updating process continuously, and we can obtain the final cumulative attention maps until the classification model converges. In the above updating process, the attention fusion strategy is responsible for preserving the discriminative regions of these intermediate attention maps to constitute more complete object regions.

Regarding the fusion strategy, we propose an effective but simple one, which is the element-wise maximum operation. It takes the maximum attention values between the attention maps A_t and the current cumulative attention maps M_{t-1}, which is formulated as follows:

$$M_t = \mathrm{AF}\big(M_{t-1}, A_t\big) = \max\big(M_{t-1}, A_t\big). \tag{5}$$

The OAA with maximum fusion strategy can effectively save the different discriminative object regions into the cumulative attention maps. As shown in Fig. 5, the cumulative attention maps generated by OAA have more entire regions than the attention maps generated by CAM [43]. We also explore the averaging fusion strategy for OAA. However, the performance drops 1.6% of the mIoU scores compared to the maximum fusion strategy. In Sec. 4.3, we perform ablation experiments to show the differences between these two fusion strategies.

It is worth mentioning that as the classification model is weak and may focus on noisy regions at the beginning of the training process, we use the predicted probability of the target classes to decide whether we accumulate the corresponding attention maps. In particular, if the classification score of the target category is higher than those of all non-target categories, we accumulate the attention map of the target category in OAA. Otherwise, we abandon this attention map to avoid noise.

3.3. Towards Integral Attention Learning

The OAA integrates the attention maps at different epochs in the training phase to produce more integral object regions. However, the weakness of OAA is that some object regions with lower attention values cannot be enhanced by the classification model itself. Taking this situation into account, we introduce a new loss function by regarding the cumulative attention maps as supervision to train an integral attention model to further improve our OAA, which is named as OAA$^+$.

To be specific, we use the cumulative attention maps as soft labels as done in [36]. Each attention value is viewed as the probability of the location belonging to the corresponding target class. We adopt the classification network shown in Fig. 2 without the global average pooling layer and classification loss as our integral attention model. Given the score map $\hat{F}$ produced by the class-aware convolutional layer, the probability of location j being some category c can be denoted by $q_j^c = \sigma(\hat{F}_j^c)$, where σ is the sigmoid function. Thus, the multi-label cross-entropy loss for class c used in

[1]Here, we omit the class c for convenience.

[36] can be written as:

$$-\frac{1}{|N|}\sum_{j\in N}\left(p_j^c\log(q_j^c)+(1-p_j^c)\log(1-q_j^c)\right), \quad (6)$$

where p_j^c denotes the values in the normalized cumulative attention maps. After optimization, the enhanced attention maps can be obtained directly from the class-aware convolutional layer. However, with the above multi-label cross entropy loss function, the produced attention maps tend to cover the semantic object regions partially. The reason is that the loss function in Eq. (6) prefers classifying pixels with low class-specific attention values ($p_j^c < 1 - p_j^c$) to be the background for category c.

In consideration of the above discussion, we propose an improved hybrid loss. Given the cumulative attention map ranging from 0 to 1 for class c, we firstly divide it to soft enhance regions N_+^c and soft constraint regions N_-^c, where N_-^c includes pixels with $p_j^c = 0$ and N_+^c contains other pixels. For pixel set N_+^c, we remove the last term of Eq. (6) in order to further promote the attention regions but not suppress the regions with low attention values. Formally, we have the loss function for N_+^c as

$$\mathcal{L}_+^c = -\frac{1}{|N_+^c|}\sum_{j\in N_+^c} p_j^c\log(q_j^c). \quad (7)$$

As only image-level labels are given here, the attention regions in the cumulative attention maps often contain non-target pixels because of the irregular shapes of semantic objects. Therefore, in Eq. (7), we use p_j^c as the ground-truth label instead of 1 such that lower attention values in the cumulative attention maps over non-semantic areas have nearly no negative effect on the network. For N_-^c where $p_j^c = 0$, the loss function in Eq. (6) collapses to the following form:

$$\mathcal{L}_-^c = -\frac{1}{|N_-^c|}\sum_{j\in N_-^c}\log(1-q_j^c). \quad (8)$$

As a result, the total hybrid loss function for our integral attention model can be computed by:

$$\mathcal{L} = \sum_{c\in C}(\mathcal{L}_+^c + \mathcal{L}_-^c). \quad (9)$$

In this way, the lower values in soft enhanced regions also contribute to optimization according to the loss function in Eq. (7). Eq. (8) constrains the excess expansion of attention areas to the background.

Based on the proposed loss function, we can train an integral attention model to further strengthen the lower attention values of target object regions. At the inference time, the improved attention maps can be directly obtained from the class-aware convolutional layer of the integral attention model. Additionally, Fig. 5 shows some visual results of our attention maps, and more quantitative analysis is conducted in Sec. 4.3.

4. Experiments

In order to demonstrate the effectiveness of our approach, we apply our attention maps produced by OAA and OAA$^+$ as heuristic cues to the weakly-supervised semantic segmentation task. We use the attention maps to extract object cues and saliency maps [12] to extract background cues. These cues are then utilized to generate the pseudo segmentation annotations. We assign the category tag corresponding to the maximum value to the pixels in proxy segmentation labels. All the conflicted pixels are ignored for training. The proxy ground-truths generated from the above method are used to train segmentation models. In the following subsections, we provide a series of ablation studies and compare our approach with the previous state-of-the-art approaches.

4.1. Dataset and Settings.

Dataset and Evaluation Metrics We evaluate our approach on the PASCAL VOC 2012 segmentation benchmark [8], which contains 20 semantic categories and the background. As done in most previous work, we also use the augmented training set [9] for model training. Therefore, we have 10,582 training images in total. During the test phase, we compare our approach with previous methods on both the validation and test sets in terms of the mean intersection-over-union (mIoU) evaluation metric. Because the segmentation annotations for the test set are not publicly available, we submit the predicted results to the official PASCAL VOC evaluation server to obtain the scores.

Network Settings. For the classification network, the hyper-parameters are set as follows: mini-batch size (5), weight decay (0.0002), and momentum (0.9). The initial learning rate is set to 1e-3, which is divided by 10 after 20000 iterations. We run the classification network for 30000 iterations in total. We use the classification network without the global average pooling layer and classification loss as our integral attention model. The hyper-parameters of the integral attention model is the same as that of the classification network. We use the DeepLab-LargeFOV model [5] as done in most previous work as our segmentation network. The segmentation network is trained with a mini-batch of 10 images and terminated at $15,000$ iterations. All the other hyper-parameters are the same as [5]. We report results based on both VGG16 [32] and ResNet-101 [10] backbones.

4.2. Comparisons to the State-of-the-arts

In this subsection, we compare our approach with previous weakly-supervised semantic segmentation methods relying on only image-level labels. Tab. 1 lists all the results of these approaches and ours on the validation and test sets. It can be easily observed that the mIoU scores of our approach improve all the previous state-of-the-art methods, no matter which backbone is used. Among the previous state-

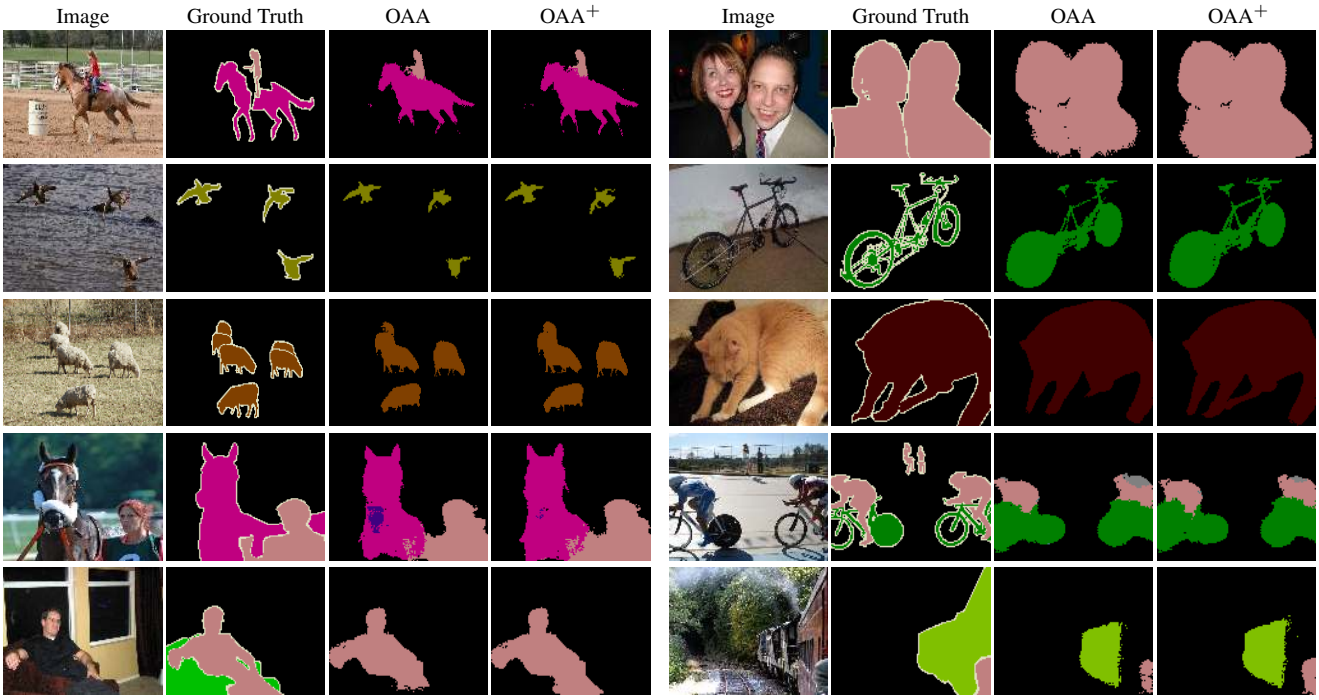

Figure 4. Qualitative segmentation results on the PASCAL VOC 2012 validation set using attention maps generated by our OAA and OAA$^+$, respectively. We also show several failure cases on the bottom row.

of-the-art methods, MIL [26] and WebS-i2 [16] use more training images (700K and 19K, respectively). Furthermore, Hong *et al*. [11] utilizes rich information of the temporal dynamics provided by additional video data, which helps easily find out the integral semantic objects from video data. Although only 10K images are used, the results of our OAA$^+$ approach improve the above three approaches on the validation set by 21.1%, 9.7% and 5.0%, respectively. This fact well demonstrates that the attention maps produced by our integral attention model can effectively detect more integral semantic regions towards all parts of the target objects.

Comparing to AE-PSL [35], our OAA achieves a better mIoU score (61.6% *v.s.* 55.0%) with no need to train multiple classification models. Furthermore, GAIN [19] adopts a self-guidance erasing strategy in an end-to-end manner but our segmentation results improve GAIN by more than 7% mIOU score (63.1% *v.s.* 55.3%). The comparisons to those erasing-based methods reveal that collecting the intermediate attention maps is more effective. In [37], Wei *et al*. exploited the power of dilated convolutions to discover integral objects. However, it usually introduces some irrelevant pixels because the convolutions with large dilation rates often focus on the outside of the target regions. Differently, our approach does not utilize convolutions with large dilation rates and hence can weaken the effects of irreverent pixels. As shown in Tab. 1, our approach improves the method of [37] by nearly 2% on both the validation set and the test set. Additionally, we also show the segmentation results based on ResNet [10] backbone. Obviously, our proposed approach achieves the best result on the PASCAL VOC 2012 segmentation benchmark.

4.3. Ablation Analysis

In this section, we perform a series of ablation experiments and give detailed analysis to demonstrate the effectiveness of the proposed strategies. Furthermore, we demonstrate how the produced attention maps can benefit the semantic segmentation task. Note that we use the VGGNet version DeepLab-LargeFOV model in this subsection.

Accumulation Strategies. The attention fusion strategy is used in OAA to accumulate the discovered discriminative regions in the intermediate attention maps at different epochs. In addition to the maximum fusion strategy, we also investigate an average fusion strategy, which can be formulated as:

$$M_t = \frac{1}{t}\big((t-1)M_{t-1} + A_t\big). \tag{10}$$

As shown in Tab. 2, using attentions by CAM [43] without OAA gives a mIoU score of 53.9% on the validation set. When adding OAA with the average fusion strategy, the result can be improved to 57.0%. When replacing the average fusion strategy with the maximum fusion strategy, we have a mIoU score of 58.6%, which greatly improves the results based on CAM [43]. In addition, we observe that OAA with the maximum fusion strategy is more effective than that with the average fusion strategy. This is because the averaging

Methods	Supervision	Val	Test
Backbone: VGGNet [32]			
CCNN [25]	10K	35.3%	-
EM-Adapt [24]	10K	38.2%	39.6%
MIL [26]	700K	42.0%	-
DCSM [30]	10K	44.1%	45.1%
SEC [18]	10K	50.7%	51.7%
AugFeed [27]	10K	54.3%	55.5%
STC [36]	50K	49.8%	51.2%
Roy et al. [28]	10K	52.8%	53.7%
Oh et al. [23]	10K	55.7%	56.7%
AE-PSL [35]	10K	55.0%	55.7%
Hong et al. [11]	970K	58.1%	58.7%
WebS-i2 [16]	19K	53.4%	55.3%
DCSP [3]	10K	58.6%	59.2%
TPL [17]	10K	53.1%	53.8%
GAIN [19]	10K	55.3%	56.8%
DSRG [15]	10K	59.0%	60.4%
MCOF [34]	10K	56.2%	57.6%
Ahn et al [1]	10K	58.4%	60.5%
Wei et al [37]	10K	60.4%	60.8%
SeeNet [14]	10K	61.1%	60.7%
OAA (Ours)	10K	61.6%	61.9%
OAA$^+$ (Ours)	10K	63.1%	62.8%
Backbone: ResNet [10]			
DCSP [3]	10K	60.8%	61.9%
DSRG [15]	10K	61.4%	63.2%
MCOF [34]	10K	60.3%	61.2%
Ahn et al [1]	10K	61.7%	63.7%
SeeNet [14]	10K	63.1%	62.8%
OAA (Ours)	10K	63.9%	65.6%
OAA$^+$ (Ours)	10K	65.2%	66.4%

Table 1. Quantitative comparisons to previous state-of-the-art approaches on both the validation and test sets. OAA$^+$ denotes that the attention maps are generated from the integral attention model described in Sec. 3.3.

fusion strategy averages all the attention values across the intermediate attention maps, which decreases the attention values in the final cumulative attention maps. Therefore, in the following, we view the maximum fusion strategy as our default fusion strategy for OAA. Note that the goal of this paper is to demonstrate the effectiveness of OAA and hence we simply choose the element-wise maximum fusion strategy for OAA. Designing more complicated fusion strategy is beyond the scope of this paper but we encourage readers to further explore more effective ones.

Loss Function in OAA$^+$. As stated in Sec. 3.3, the cumulative attention maps are then used as soft labels to train the integral attention model to produce attention maps with more integral and accurate object regions. In Tab. 2, we show quantitative results using different loss functions. It can be observed that the performance is improved by 8.4% when replacing the standard multi-label cross-entropy loss (MCE) [36] with the proposed hybrid loss (HL). When applying the multi-label cross entropy loss, the output attention maps always cover small object regions. On the contrary, the proposed hybrid loss can further improve the quality of the cumulative attention maps by our OAA.

No.	AVE	MAX	MCE	HL	mIoU (val)
1					53.9%
2	✓				57.0%
3		✓			58.6%
4		✓	✓		51.2%
5		✓		✓	59.6%

Table 2. Comparisons of mIoU scores on the PASCAL VOC 2012 validation set when using different settings. **AVE:** OAA with the average fusion strategy. **MAX:** OAA with the maximum fusion strategy. **MCE:** OAA$^+$ using the multi-label cross entropy loss in Eq. (6). **HL:** OAA$^+$ using the proposed hybrid loss in Eq. (9).

Results with Different Strategies. Other than visual comparisons, we also perform a series of ablation experiments on the PASCAL VOC 2012 dataset. As shown in Tab. 2, we show that the mIoU scores of using attention maps with different strategies for training segmentation networks. In the third and the last rows of Tab. 2, it can be seen that using OAA$^+$ can further improve the results by OAA by 1.0% on the validation set, which indicates our integral attention model with the proposed loss function can help further improve the quality of the cumulative attention maps.

Visual Comparisons. In this paragraph, we show some qualitative results on the PASCAL VOC 2012 dataset [8] and give the corresponding attention maps produced by CAM [43], OAA, and OAA$^+$, respectively, for visual comparisons. As shown in Fig. 5, the images include different kinds of scenes, such as images with objects of different scales, crowded objects, and multiple categories. From all shown examples, our cumulative attention maps can discover nearly complete target objects at different scales, when comparing to the attention maps produced by CAM [43]. On the fifth row, the images with multiple objects are shown. It can be found that in this case our cumulative attention maps can still cover most of the semantic regions. On the last row, we show some examples containing multiple classes. Obviously, our cumulative attention maps can successfully distinguish different classes and detect the target objects densely. In addition, the attention maps produced by OAA$^+$ can discover more integral object regions than the cumulative attention maps from OAA. Additionally, we also show some segmentation results in Fig. 4.

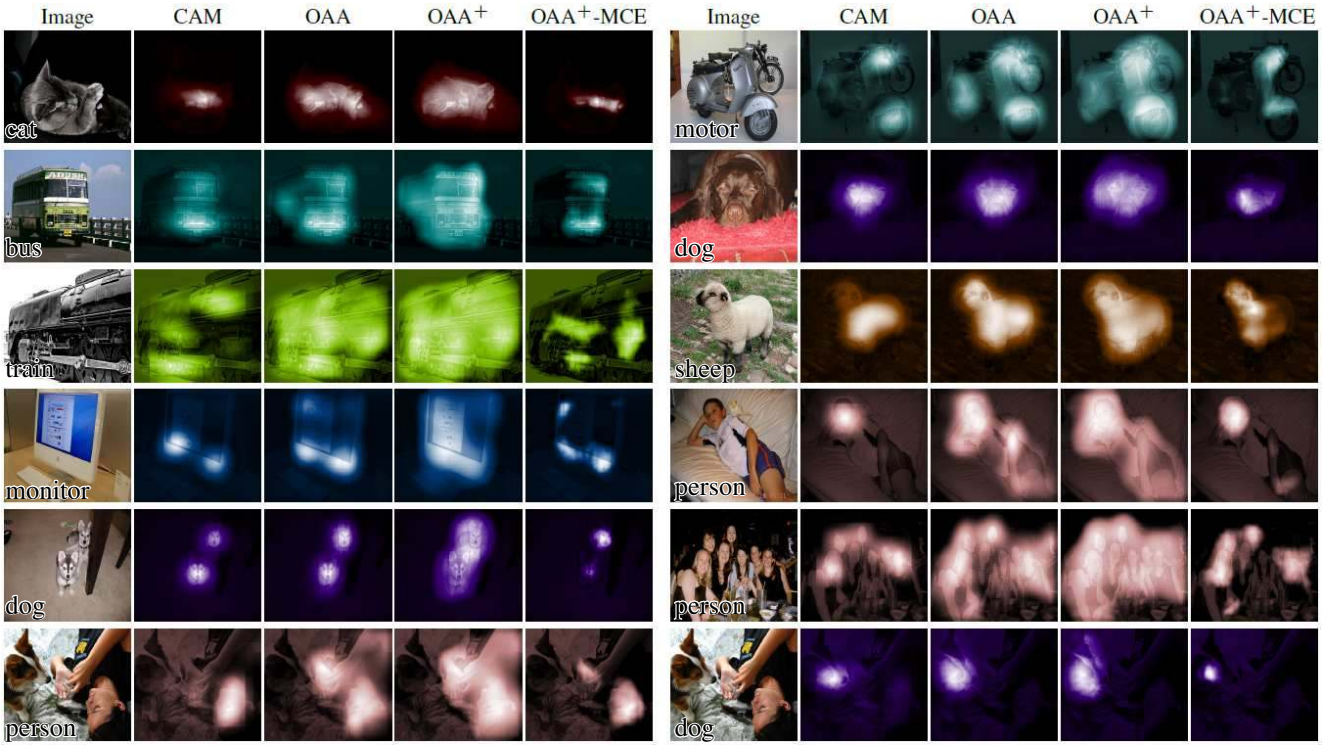

Figure 5. Visual comparisons among different attention maps produced by CAM [43], OAA, OAA$^+$ and OAA$^+$-MCE. OAA$^+$ and OAA$^+$-MCE denote the integral attention model learned with the proposed hybrid loss in Eq. (9) and the multi-label cross entropy loss in Eq. (6) respectively.

No.	#Training Images	Proportion	mIoU(val)
1	2,116	20%	54.6%
2	5,291	50%	57.3%
3	8,466	80%	58.9%
4	10,582	100%	59.6%

Table 3. Comparisons of mIoU scores on PASCAL VOC 2012 validation set when using different number of training images. Note that images are selected randomly. **Proportion:** the percentage of the images used for training. **#Training Images:** the number of training images.

Number of Training Images. To further investigate the quality of the attention maps, we attempt to use different numbers of training images to train the segmentation network. We use the attention maps produced by OAA$^+$ to produce the proxy segmentation annotations. As shown in Tab. 3, the mIoU scores are improved gradually as more images are used for training. More interestingly, when using only 2116 training images, our segmentation network can still achieve a performance score of 54.6%, which is better than the segmentation results based on CAM [43]. This indirectly suggests that our attention maps are with high quality and facilitate the segmentation task.

5. Conclusion

In this paper, we explore a simple but effective framework called OAA to discover more integral object regions. We maintain a series of cumulative attention maps to preserve the different discriminative regions in attention maps generated by the classification network during training stages. Additionally, we utilize the cumulative attention maps as soft labels to train an integral attention model to enhance the attention maps by OAA. Our approach is easy to follow and can be simply plugged into any classification networks to discover the target object regions holistically. Thorough experiments show that when applying our attention maps to the weakly-supervised segmentation task, our segmentation network works better than the previous state-of-the-arts. In the future, we plan to conduct experiments on large-scale datasets, such as MS COCO [21] and ImageNet [7].

Acknowledgment This research is supported by NSFC (61572264, 61620106008), the national youth talent support program, and Tianjin Natural Science Foundation (17JCJQJC43700, 18ZXZNGX00110). Yunchao Wei is partly supported by IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) and ARC DECRA DE190101315.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 2, 7

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 1

[3] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017. 7

[4] 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. 1

[5] 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 TPAMI*, 2017. 5

[6] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 2

[7] Jia Deng. A large-scale hierarchical image database. In *CVPR*, 2009. 8

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 2, 5, 7

[9] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6, 7

[11] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017. 6, 7

[12] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019. 2, 5

[13] Qibin Hou, Puneet Kumar Dokania, Daniela Massiceti, Yunchao Wei, Ming-Ming Cheng, and Philip Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. *EMMCVPR*, 2017. 2

[14] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018. 2, 7

[15] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018. 2, 7

[16] Bin Jin, Maria V Ortiz Segovia, and Sabine Susstrunk. Webly supervised semantic segmentation. In *CVPR*, 2017. 6, 7

[17] Dahun Kim, Donggeun Yoo, In So Kweon, et al. Two-phase learning for weakly supervised object localization. In *ICCV*, 2017. 7

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*. Springer, 2016. 2, 3, 7

[19] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018. 2, 3, 6, 7

[20] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In *CVPR*, 2017. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 8

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[23] 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. 7

[24] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015. 7

[25] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 7

[26] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 1, 6, 7

[27] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016. 1, 7

[28] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*, 2017. 7

[29] 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. 1, 2

[30] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 7

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3, 5, 7

[33] Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Xiaowei Hu, and Nanning Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 123(2):251–268, 2017. 2

[34] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 2, 7

[35] 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. 1, 3, 6, 7

[36] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 39(11):2314–2320, 2017. 4, 5, 7

[37] 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. 2, 3, 6, 7

[38] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 1

[39] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 2

[40] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 1, 2, 3

[41] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, 2018. 2

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

[43] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 2, 3, 4, 6, 7, 8