Online Attention Accumulation for Weakly Supervised Semantic Segmentation

Peng-Tao Jiang*, Ling-Hao Han*, Qibin Hou, Ming-Ming Cheng, Yunchao Wei

Abstract—Object attention maps generated by image classifiers are usually used as priors for weakly supervised semantic segmentation. However, attention maps usually locate the most discriminative object parts. The lack of integral object localization maps heavily limits the performance of weakly supervised segmentation approaches. This paper attempts to investigate a novel way to identify entire object regions in a weakly supervised manner. We observe that image classifiers' attention maps at different training phases may focus on different parts of the target objects. Based on this observation, we propose an online attention accumulation (OAA) strategy that utilizes the attention maps at different training phases to obtain more integral object regions. Specifically, we maintain a cumulative attention map for each target category in each training image and utilize it to record the discovered object regions at different training phases. Albeit OAA can effectively mine more object regions for most images, for some training images, the range of the attention movement is not large, limiting the generation of integral object attention regions. To overcome this problem, we propose incorporating an attention drop layer into the online attention accumulation process to enlarge the range of attention movement during training explicitly. Our method (OAA) can be plugged into any classification network and progressively accumulate the discriminative regions into cumulative attention maps as the training process goes. Additionally, we also explore utilizing the final cumulative attention maps to serve as the pixel-level supervision, which can further assist the network in discovering more integral object regions. 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 67.2% on the test set.

Index Terms—Weakly supervised semantic segmentation, attention maps, online attention accumulation, attention drop layer, pixel-level supervision.

1 Introduction

Benefiting from the advanced convolutional neural network (CNN) architectures [4]–[6], fully-supervised semantic segmentation approaches, such as [7]–[13], have made remarkable performance recently. However, these CNN-based segmentation approaches heavily rely on large-scale training data with pixel-level annotations. Constructing a pixel-accurate segmentation dataset is fairly expensive, as it requires considerable human effort and time cost. To economize on human labor, researchers propose to learn semantic segmentation with weak supervision. The common weak supervisions for segmentation include bounding boxes [14], scribbles [15], points [16], and image-level labels [17]. Among these weak supervisions, image-level labels are more easily obtained than other weak supervisions [18] and hence are widely adopted for semantic segmentation.

The difficulty of learning semantic segmentation with image-level labels is that these labels can only tell whether the target objects exist in an image but don't provide any object location and shape information. Recently, many works [19]–[24] attempt to learn semantic segmentation with image-level labels based on the class activation maps [25]. Because of the ability to discover pixel-level information, class activation maps have been widely used in the weakly supervised semantic segmentation task for generating initial class-specific object regions. However, the original class activation maps often focus on small parts of the semantic objects, limiting the capability of learning richer pixel-level semantic knowledge for segmentation networks. Later methods consider leveraging different adversarial erasing strategies [26]–[28] to enlarge the discriminative regions. Unfortunately, as the training process continues, the discriminative regions expand anticipated, making some undesired background stuff predicted as semantic regions. In [22], dilated convolution is leveraged for integral attention generation. However, a similar problem occurs when the dilation rates increase.

Although the above approaches are different from each other, one common point shared by them is that they all utilize the final classification models to generate attention maps. In this paper, we start from a new perspective and consider the way of attention generation by considering the classification models' training process. We observe that the discriminative regions in attention maps generated at different training phases constantly shift to different parts of the semantic objects before the classification network reaches convergence. The main reasons can be briefly summarized as follows: (i) First, a powerful classification network usually seeks robust common patterns for a specific category so that all the images from this category can be well recognized. As pointed by [29], the network prioritizes learning simple patterns first. 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 the discriminative regions until the

- *P.T. Jiang, L.H. Han, and M.M. Cheng are with TKLNDST, College of Computer Science, Nankai University, Tianjin, China, 300350. M.M. Cheng is the corresponding author (cmm@nankai.edu.cn). * denotes equal contribution.*
- *Q. Hou is with the Department of Electrical and Computer Engineering, NUS, Singapore, 119077.*
- *Y. Wei is with the Institute of Information Science, Beijing Jiaotong University, Beijing, China, 10044.*
- *A preliminary version of this work appeared at ICCV [1]. Both PyTorch [2] and Jittor [3] versions of the source code are publicly available via our project page: http://mmcheng.net/oaa/.*

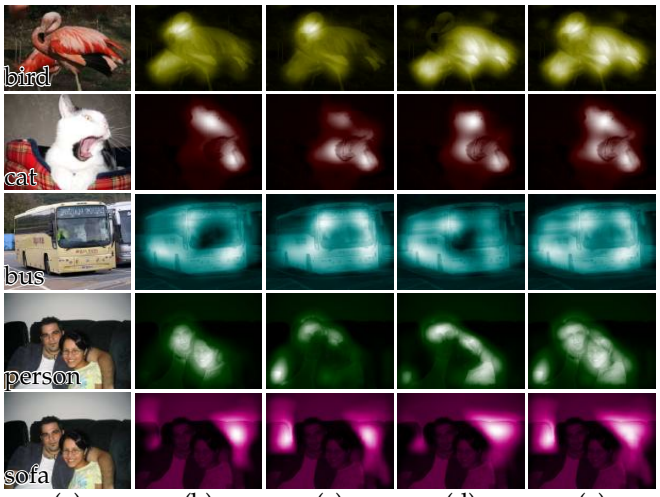


Fig. 1. Observation of our proposed approach. (a) Source images; (b-d) Intermediate attention maps produced by a classification network at different training stages; (e) Cumulative attention maps produced by combining attention maps in (b), (c), and (d) through a simple element-wise maximum operation. 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 the semantic regions compared with (b), (c), and (d). Best viewed in color.

network reaches convergence. Zeiler et al. [30] have also shown that deep layers of the network usually converge using a considerable number of epochs. Thus, with the constant change of the parameters in deep layers, the attention generated from deep layers continuously shifts during the training phase. (ii) Second, during training, the current classification model's attention maps are mostly influenced by the previous input images. Therefore, images with different content and the training images' input order will lead to the variation of the discriminative regions during training. Fig. 1(b-d) gives a clear illustration of this phenomenon. Moreover, we also observe that the discriminative regions discovered at different training phases are often complementary, as shown in Fig. 1(e). This fact motivates us to record the intermediate attention maps for detecting complete semantic objects with only image-level labels.

Based on the above observation, we introduce a simple yet effective approach for attention generation, capable of taking attention maps from different training phases into account. Specifically, we present an online attention accumulation (OAA) strategy. A cumulative attention map for each category in each image is maintained to sequentially accumulate the discriminative regions generated by the attention maps at different training phases. The complementarity of the intermediate attention maps enables discovering integral object objects to be possible. Despite the relatively complete attention regions by OAA compared with CAM [25], some attention values in object regions are still not strong enough. To improve this situation, in our original conference version, we 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. The new attention model advances the cumulative attention maps and can generate more integral object regions.

In this paper, we also improve the cumulative attention maps by rethinking and improving the OAA strategy. We observe that for some training images, the range of attention movement is not large. As shown in Fig. 2, attention of intermediate attention maps always moves slightly on the bird's head and the left sheep, respectively. Through the combination of Fig. 2(b-d), we still cannot find the body of the bird and the right sheep. Early works [26], [27], [31] erase the strong attention regions to force the final classification model to locate other parts of target objects. Inspired by the erasing strategy, we further propose integrating an attention drop layer into our online attention accumulation process. The goal is to enlarge the range of attention movement during training to discover new object regions. Specifically, by a certain probability, the attention drop layer occludes the regions with strong activations on the input images and feeds them into the classification network. Then OAA accumulates the attention maps with new object regions. By utilizing the simple attention drop layer, OAA can locate more integral object regions without an additional training process. Experiments demonstrate that the proposed attention drop layer further improves the segmentation performance from 66.4% (our conference version) to 67.2% in terms of mean IoU on the test set of the PASCAL VOC 2012 [32]. Additionally, we also conduct experiments to help readers better understand the attention evolution process during training.

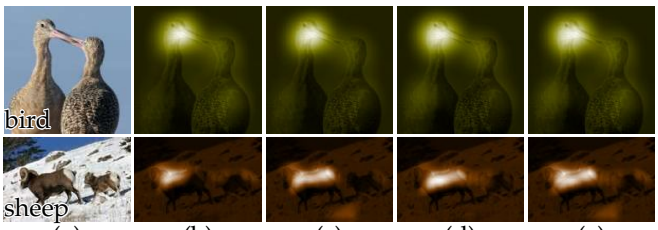


Fig. 2. The motivation of adding an attention drop layer. (a) Source images; (b-d) Intermediate attention maps at different training stages; (e) Cumulative attention maps produced by combining attention maps in (b-d) through a simple element-wise maximum operation. We can see that the intermediate attention maps focus on the same object regions, causing the cumulative attention maps hard to find new object regions.

2 Related Work

This section briefly reviews the history of visual attention methods and describes the weakly supervised semantic segmentation methods strongly related to our work.

2.1 Visual Attention for Localization

To date, some outstanding works have been proposed, which utilize the classification models to locate reliable object regions. As an early attempt, CAM [25] showed that the global average pooling (GAP) layer could be used to generate class activation maps, which locate the most discriminative object regions. Later, based on CAM, Grad-CAM [33] proposed 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 class attention maps. Grad-CAM++ [34] further improved the localization ability of Grad-CAM by considering the importance of each pixel in an activation map. Score-CAM [35] computed the

weight of each activation map through its forward passing score instead of the gradient. Besides, Fong et al. [36], [37] added perturbations to the input to generate attention maps. NormGrad [38] inserted a virtual identity layer to generate attention maps for different layers. 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 signals downwards in the network via a probabilistic Winner-Take-All process.

Recently, different from the above methods for localizing the most discriminative object regions, some works [19], [22], [27], [40], [41] aim to produce attention maps by localizing large and integral relevant regions of the semantic objects, which are beneficial for the weakly supervised learning tasks. These methods utilize the final classification models to generate attention, which is different from our method that utilizes the intermediate states of classification models during training. Besides visual attention, some researchers [22], [42], [43] also found that bottom-up salient object cues [44]–[47] are very useful for extracting background cues and object shape information.

Our OAA iteratively accumulates the intermediate attention maps of classification models to update the cumulative attention maps. Such an iterative attention fusion strategy can be regarded as a generalization of recurrent attention [48]. Ba et al. [48] proposed a deep recurrent attention network to locate and recognize multiple objects in images. At each step, the network processes an input patch (called glimpse) and uses the information to update the internal representations and output the location of the next glimpse. Unlike they utilize recurrent attention to update the internal representation of the recurrent network, our iterative attention fusion strategy doesn't update the parameters of the classification network. Our OAA only utilizes the attention maps during training to update the cumulative attention map. Wang et al. [49], [50] incorporated multiple attention steps to encourage mutual interaction between different branches. Unlike they utilize multiple attention steps to process different information, our OAA incorporates multiple attention maps into the cumulative attention map to generate integral object regions.

2.2 Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation has also experienced great progress as a variety of methods were proposed. One kind of early methods [17], [51], [52] utilized the multiple instance learning (MIL) framework to learn weakly supervised semantic segmentation. Pathak et al. [23] imposed constraints on the output labeling of CNN. The ExpectationMaximization Algorithm is applied by Papandreou et al. [53] to the weakly supervised semantic segmentation. The performance of the above methods is unsatisfactory due to the lack of pixel-level information.

Later, with the help of the attention maps' localization ability, many methods utilize attention maps to extract initial object regions of target objects and refine them during training. Typically, SEC [24] introduced three loss functions called seeding, expansion, and boundary constrain losses to expand the initial seeds and train the segmentation model carefully. However, the performance of SEC is limited in that the object-related seeds from CAM [25] only cover small and sparse semantic regions. DSRG [21] also utilized the seeded region growing strategy. They progressively increased the quality of segmentation labels during training.

More recently, many works are two-folds, *i.e.*, generate accurate pseudo segmentation labels first, then use the pseudo labels to train existed segmentation models. One key to generating high-quality segmentation labels is to locate target object regions accurately and completely. AffinityNet [20] proposed to learn semantic affinities among adjacent pixels and then generate high-quality segmentation labels by transferring the semantics of known pixels to their adjacent unknown pixels. Wei et al. [22] mined more non-discriminative object regions by revisiting the dilated convolution with different dilation rates. Wang et al. [54] first applied a transformation constraint to revise attention maps and then utilize AffinityNet [20] to refine the attention maps further. Chang et al. [55] utilized the sub-category information to force the network to pay attention to the integral object regions. Sun et al. [56] improved the localization ability of attention maps by leveraging the cross-image semantic relations. Besides, Shimoda et al. [57] proposed a refined process to improve the quality of pseudo segmentation masks.

Moreover, researchers have utilized an erasing-based strategy to mine integral target object regions based on classification networks. In [26], Wei *et al.* proposed an approach that used an adversarial erasing (AE-PSL) strategy to mine different regions of the objects progressively. However, AE-PSL procedures are complicated, which requires repetitive training procedures and learns multiple classification models to obtain different object regions. GAIN [40] improved the adversarial erasing strategy by using attention maps to provide a self-guidance that forced the network to focus on the objects holistically. ACoL [27] introduced another classification head to the network and performed the adversarial erasing strategy in an end-to-end manner.

Our work also falls into this category that generates pseudo labels first. Unlike these methods only generating attention maps from the final classification models, we start from a new perspective and attempt to mine more object regions by considering the intermediate states of training. We also utilize the erasing-based strategy to the classification network's training process to increase the range of attention movement during training. Experiments show that the online attention accumulation process is effective in obtaining integral object regions.

3 Methodology

We explain the proposed methods in detail and then delve into each component's specifics in our framework. The pipeline of our approach is shown schematically in Fig. 3, Fig. 4, and Fig. 6.

3.1 Attention Generation

To obtain the object regions, we employ CAM [25] as the default attention map generator. Specifically, we utilize the class-aware feature maps outputted by the last convolutional layer to generate attention maps [25], [27], which could be generated during the training stage with almost no effort.

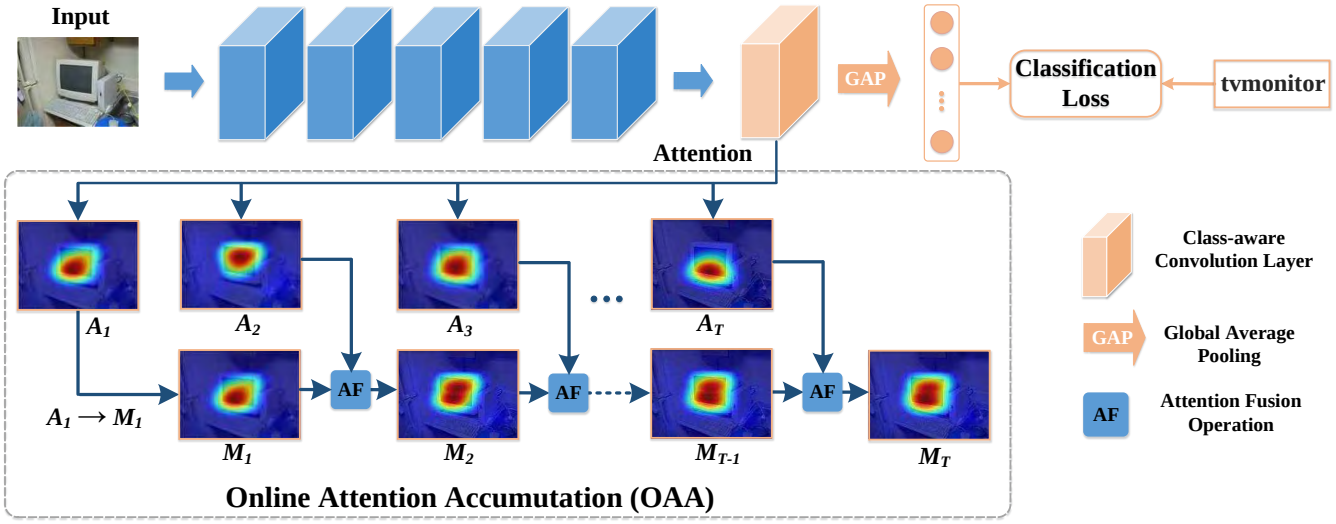


Fig. 3. Illustration of our online attention accumulation (OAA) process. The attention maps are generated online from the class-aware convolutional layer. Our OAA utilizes these discriminative regions of attention maps at the different training phases and integrates them into the cumulative attention maps with a simple attention fusion strategy progressively.

Following most of the previous works [22], [27], we also take the VGG-16 [4] as the backbone model. Three convolutional layers are added to the backbone models' last convolutional layer for nonlinear transformation. The output is then connected to a class-aware convolutional layer with the channel number of C and the kernel size of 1×1 for capturing the attention, where C denotes the number of categories. As shown at the top of Fig. 3, attention maps are generated by the last convolutional layer's output before the global average pooling (GAP) layer. Regarding the fact that some images may have more than one category, we treat the whole training process as C binary classification problems. Therefore, the probability of the target category c could be predicted by

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

where $\sigma(\cdot)$ denotes the sigmoid function. The cross-entropy loss is applied to optimize the network.

To generate the attention map for an image I, category c, during training, we first perform a ReLU operation on the feature map F^c and then normalize it by

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

With the attention maps generated at different training stages, integral attention is gradually formed in the OAA process.

3.2 Online Attention Accumulation

According to our observation, the attention maps during different training phases usually focus on different parts of the target objects. Thus, we propose an online attention accumulation (OAA) strategy to fully use the attention maps in training. OAA accumulates the attention maps for any training image, generated at different training epochs, into a cumulative attention map. As shown in Fig. 3, given a training image I, we create a cumulative attention map M^c for target class c to record the discriminative regions discovered at each training epoch.

To be specific, when the training image I is fed into the classification network at the first epoch, the generated attention map A_1^c of the target category c is utilized to initialize the cumulative attention map M_1^c. For simplicity, we will omit the category c in the following notations. Then, when the image is inputted into 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. In the same way, the attention map A_t at the t-th epoch is used 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 these intermediate attention maps' discriminative regions to constitute more complete object regions.

Regarding the fusion strategy, we propose to use the element-wise maximum operation. It takes the maximum attention values between the attention map A_t and the current cumulative attention map M_{t-1}, which is formulated as follows:

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

With the maximum fusion strategy, the OAA can be implemented effectively and involves little training effort to fuse the different parts of objects into the cumulative attention maps. As shown in Fig. 7, the cumulative attention maps generated by OAA mine more complete object regions than the attention maps generated by CAM [25]. Besides, we conduct experiments on the OAA with the averaging fusion strategy. However, the score of the mIoU decreases by 1.6% compared with the maximum fusion strategy. In Sec. 4.3, we perform ablation experiments to show the differences between these two fusion strategies.

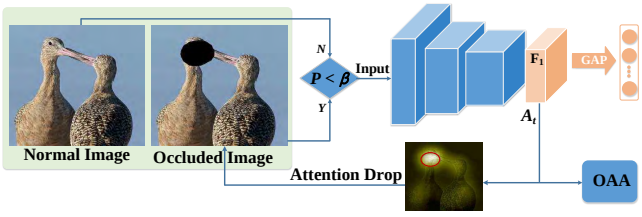


Fig. 4. Illustration of integrating an attention drop layer into our online attention accumulation (OAA) process. The red circle denotes the regions whose attention values are large than δ.

It should be noted that, at the beginning of the training process, the classification model is less accurate, and the generated attention maps may not contain the object regions. So we use the predicted classification score of the target category to decide whether to accumulate the current attention maps or not. If the target category's classification score is higher than all non-target categories, we accumulate the target category's attention map in OAA. Otherwise, this attention map is discarded to avoid noise.

3.3 Integrating an Attention Drop Layer

As mentioned in Sec. 1, we observe that some images' attention shifts slightly on the object regions during the training process. It's hard to mine not emerging object regions for these images through our online attention accumulation process. Recently, some researchers [26]–[28], [31] attempt to occlude the attention regions in input images or features to impel network attention to focus on other discriminative regions of the objects for recognition. Inspired by these works, we attempt to integrate the erasing-based strategy into our online attention accumulation process to expand the range of attention shifts during training to accumulate more undiscovered object regions and obtain integral object regions.

Based on such considerations, we propose to integrate an attention drop layer into our online accumulation process. Specifically, the attention drop layer is used according to a drop rate β that determines how frequently the attention drop layer is used. For an input image I in the t-th epoch, we have its attention map A_{t-1} from the last epoch. When utilizing the attention drop layer, it will occlude the regions in I corresponding to attention values in A_{t-1} larger than δ. The generated attention map A_t will be accumulated into the cumulative attention map M_{t-1} by Eqn. (4). The detailed algorithm is shown in the Algorithm. 1.

Due to the attention drop layer effectively expanding the range of attention movement during training, our OAA can mine more object regions. In Fig. 5, we can see the final cumulative attention maps from OAA, with the attention drop layer, cover more object regions than OAA. Moreover, the attention drop layer is easily plug-in OAA, which can achieve close performance to the attention learning strategy in Sec. 3.4, but only needs a one-fold training process. In the ablation experiments, we analyze the impact of the drop rate β and the attention threshold δ by enumerating possible values and choosing the best performance value.

Algorithm 1 Integrating an attention drop layer

Input: Input I, attention map A_{t-1}, drop layer $S(\cdot)$, drop rate β, threshold δ, classification model $f(\cdot)$
Generate a random probability: r = Rand$(0, 1)$
if r < β **then**
  Occlude input image: $I^* = S(I, A_{t-1}, \delta)$
  Input occluded image, yield $f(I^*)$
else
  Input normal image I, yield $f(I)$
end if
Obtain current attention map A_t from model $f(\cdot)$
Update cumulative attention map M_t by Eqn. (4)

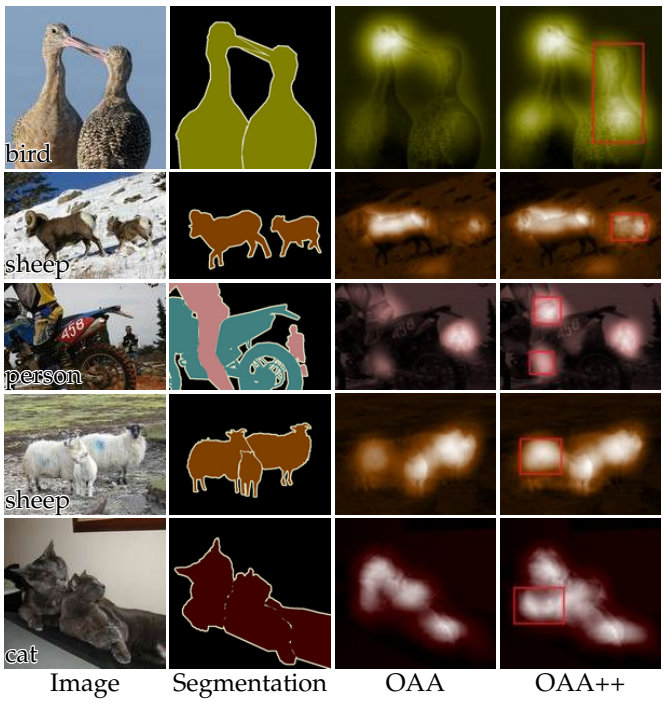


Fig. 5. Examples produced by our OAA and OAA++. **Segmentation** denotes the ground-truth segmentation masks. OAA++ newly discovers the regions in red boxes.

3.4 Towards Integral Attention Learning

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

To be specific, we use the cumulative attention maps as soft labels as done in [58]. The cumulative attention maps are normalized to [0, 1], where each value is viewed as the probability of the location belonging to the corresponding target class. We adopt the classification network shown in Fig. 3 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

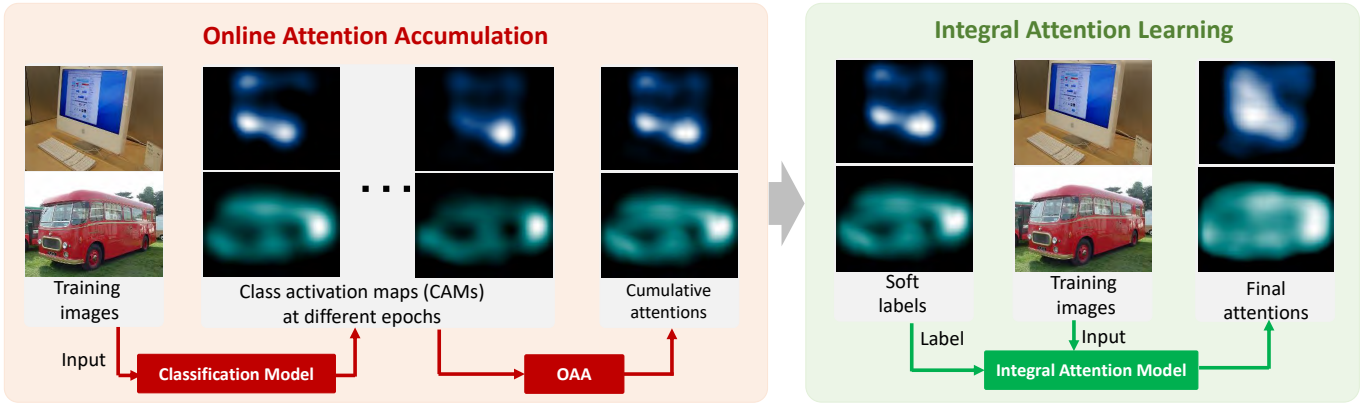


Fig. 6. The pipeline of our OAA$^+$ approach. The attention maps generated by the classification network during different training times are fused into the cumulative attention maps to mine the object regions as entire as possible. The obtained cumulative attention maps are utilized as pixel-level supervision to train the integral attention model, which further advances the attention maps' quality.

used in [58] 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 partially cover the semantic object regions. The reason is that the loss function in Eqn. (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 first divide it into 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 Eqn. (6) 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 Eqn. (7), we use p_j^c as the ground-truth label instead of 1. 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 Eqn. (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 Eqn. (7). Eqn. (8) constrains the excess expansion of attention areas to the background.

Based on the proposed loss function, we can further train an integral attention model to strengthen the target object regions' lower attention values. The improved attention maps can be directly obtained from the class-aware convolutional layer of the integral attention model at the inference time. Additionally, Fig. 7 shows some visual results of our attention maps, and more quantitative analysis is conducted in Sec. 4.3.

4 Experiments

To demonstrate our approach's effectiveness, we apply our attention maps produced by OAA as heuristic cues to the weakly supervised semantic segmentation task. These cues are then utilized to generate pseudo segmentation labels. We take a similar way as in [22] to generate pseudo segmentation labels, where we use the attention maps to extract object cues and saliency maps [44] to extract background cues. We assign the category tag corresponding to the maximum value to the pixels in pseudo segmentation labels. All the conflicted pixels are ignored for training. The proxy segmentation labels generated from the above method are then used to train segmentation models. At the inference time of segmentation models, we utilize the multi-scale test and apply DenseCRF [67] to smooth the segmentation maps. We provide a series of ablation studies in the following subsections 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 [32], which contains 20 semantic categories and the background. The images in this dataset are split into three sets: training, validation, and test set, which includes 1464, 1449, and 1456 images, respectively. As done in most previous work, we also use the augmented training set [68] for model training. Therefore, we have 10,582 training images in total. During the inference time, we compare our approach with previous state-of-the-art methods on both the validation and test sets in terms of the mean intersection-over-union

Table 1
Quantitative comparisons to previous state-of-the-art approaches. **S:** saliency maps. **IS:** instance saliency maps. **WI** and **WV**: web crawled videos and images. **P:** pixel-level supervision. OAA$^+$ denotes that the attention maps are generated from the integral attention model described in Sec. 3.4. OAA++ denotes OAA integrated with the attention drop layer. OAA++$^+$ denotes OAA$^+$ utilizes the attention maps from OAA++ as supervision.

Methods	Supervision	Val (%)	Test (%)
Backbone: VGGNet [4]			
CCNN [23]	10K	35.3	-
EM-Adapt [53]	10K	38.2	39.6
MIL [17]	S+IG	42.0	-
DCSM [59]	10K	44.1	45.1
SEC [24]	10K	50.7	51.7
AugFeed [14]	10K+S	54.3	55.5
STC [58]	10K+S+WI	49.8	51.2
Roy et al. [60]	10K	52.8	53.7
Oh et al. [61]	10K+S	55.7	56.7
AE-PSL [26]	10K+S	55.0	55.7
Hong et al. [62]	10K+WV	58.1	58.7
WebS-i2 [63]	10K+WI	53.4	55.3
DCSP [64]	10K+S	58.6	59.2
TPL [65]	10K	53.1	53.8
GAIN-SEC [40]	10K	55.3	56.8
GAIN [28]	10K	59.4	59.6
DSRG [21]	10K+S	59.0	60.4
MCOF [42]	10K+S	56.2	57.6
AffinityNet [20]	10K	58.4	60.5
MDC [22]	10K+S	60.4	60.8
AISI [66]	10K+IS	61.3	62.1
SeeNet [19]	10K+S	61.1	60.7
OAA	10K+S	61.6	62.0
OAA++	10K+S	**63.0**	**62.7**
OAA$^+$	10K+S	63.1	62.8
OAA++$^+$	10K+S	**63.7**	**63.2**
UpperBound	10K+P	70.8	71.2
Backbone: ResNet [5]			
DCSP [64]	10K+S	60.8	61.9
DSRG [21]	10K+S	61.4	63.2
MCOF [42]	10K+S	60.3	61.2
AffinityNet [20]	10K	61.7	63.7
AISI [66]	10K+IS	63.6	64.5
SeeNet [19]	10K+S	63.1	62.8
SSDD [57]	10K	64.9	65.5
SEAM [54]	10K	64.5	65.7
ScE [55]	10K	66.1	65.9
MCIS [56]	10K+S	66.2	66.9
OAA	10K+S	63.9	65.6
OAA++	10K+S	**64.9**	**66.3**
OAA$^+$	10K+S	65.2	66.4
OAA++$^+$	10K+S	**66.1**	**67.2**
UpperBound	10K+P	75.4	75.7

(mIoU) evaluation metric [9]. Because the segmentation labels for the test set are not publicly available, we submit the predicted segmentation results to the official PASCAL VOC evaluation server to obtain the mIoU scores.

Network Settings. Our method is implemented based on the Caffe library [69]. 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 are the same as that of the classification network. For the classification network integrating an attention drop layer, we also run it for 30000 iterations. We use the DeepLab-LargeFOV model [70] 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 [70]. We report results based on both VGG-16 [4] and ResNet-101 [5] backbones, respectively.

4.2 Comparisons to the State-of-the-arts

This subsection compares our approach with previous weakly supervised semantic segmentation methods only relying on image-level labels. Tab. 1 lists all the segmentation results of these approaches and ours on the validation and test sets. It can be easily observed that our approach achieves competitive results with all the previous state-of-the-art methods, no matter which backbone is used. Among the previous state-of-the-art methods, MIL [17], STC [58], and WebS-i2 [63] utilize more additional images. Furthermore, Hong *et al.* [62] utilize rich information of the temporal dynamics provided by additional video data, which easily finds the integral semantic objects from video data. Although only 10K images are used, the results of our OAA++$^+$ improve the above four approaches on the validation set by 21.7%, 13.9%, 10.3%, and 5.6%, respectively. This fact demonstrates that the attention maps produced by our integral attention model effectively detect more integral semantic regions towards all parts of the target objects, which can benefit the quality of pseudo segmentation labels.

AE-PSL [26] train multiple classification models and carries out several erasing and mining steps to constitute final attention maps. Compared with AE-PSL, our OAA++ achieves a better mIoU score (63.0% *v.s.* 55.0%) with no need to train multiple classification models. Furthermore, GAIN [28] adopts a self-guidance erasing strategy in an end-to-end manner, but our segmentation results improve GAIN by more than 4% mIoU score (63.7% *v.s.* 59.4%). In [22], Wei *et al.* exploit 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 can weaken the effects of irreverent pixels.

As shown in Tab. 1, our approach improves MDC [22] by about 3% on both the validation and test sets. Compared with the methods (AugFeed [14], Oh *et al.* [61], AE-PSL [26], DCSP [64], DSRG [21], MCOF [42], MDC [22], AISI [66] and SeeNet [19]) using saliency maps for background cues, our OAA outperforms them by a large margin. The comparisons to those erasing-based methods [19], [26], [28], [40] reveal that collecting the intermediate attention maps is more effective. The proposed attention drop layer is also based on the erasing strategy. It can further improve our

Table 2
Comparison of the weakly supervised semantic segmentation methods on PASCAL VOC 2012 validation set. The top three results are highlighted in red, green, and blue, respectively.

Methods	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
CCNN [23]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	40.7	30.4	36.3	22.2	38.8	36.9	35.3
DCSM [59]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
SEC [24]	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 3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
STC [58]	84.5	68.0	19.5	60.5	42.5	44.8	68.4	64.0	64.8	14.5	52.0	22.8	58.0	55.3	57.8	60.5	40.6	56.7	23.0	57.1	31.2	49.8
TPL [65]	82.8	62.2	23.1	65.8	21.1	43.1	71.1	66.2	76.1	21.3	59.6	35.1	70.2	58.8	62.3	66.1	35.8	69.9	33.4	45.9	45.6	53.1
MCOF [42]	85.8	74.1	23.6	66.4	36.6	62.0	75.5	68.5	78.2	18.8	64.6	29.6	72.5	61.6	63.1	55.5	37.7	65.8	32.4	68.4	39.9	56.2
GAIN [28]	87.6	76.7	33.9	74.5	58.5	61.7	75.9	72.9	78.6	18.8	70.8	14.1	68.7	69.6	69.5	71.3	41.5	66.5	16.4	70.2	48.7	59.4
DSRG [21]	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
AffinityNet [20]	87.2	57.4	25.6	69.8	45.7	53.3	76.6	70.4	74.1	28.3	63.2	44.8	75.6	66.1	65.1	71.1	40.5	66.7	37.2	58.4	49.1	58.4
MDC [22]	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
OAA	89.7	84.3	35.6	77.4	60.4	65.6	75.4	72.2	80.4	17.2	69.3	21.3	73.4	67.4	72.5	72.9	37.4	73.7	28.4	64.4	55.2	61.6
OAA$^+$	90.0	84.1	34.7	77.6	62.7	66.4	80.5	74.6	82.4	18.8	73.0	22.7	76.8	70.8	72.7	74.8	37.7	73.4	28.4	68.6	55.3	63.1
OAA++	90.1	83.8	35.3	77.3	60.6	66.9	79.0	75.1	82.4	17.7	72.5	22.7	77.1	71.5	72.1	73.8	39.9	72.2	29.8	68.8	54.7	63.0
OAA++$^+$	90.2	85.6	36.0	75.6	62.0	66.6	82.5	73.6	83.9	18.7	75.3	18.7	77.9	73.3	73.0	75.4	40.0	76.4	29.7	68.9	54.6	63.7
Upperbound	92.4	82.8	35.6	82.1	64.5	72.8	88.0	81.0	85.3	32.5	79.0	55.8	80.4	77.9	74.3	80.0	52.2	79.9	44.7	79.9	65.2	70.8

Table 3
Comparison of the weakly supervised semantic segmentation methods on PASCAL VOC 2012 test set. The top three results are highlighted in red, green, and blue, respectively.

Methods	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
DCSM [59]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
SEC [24]	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
STC [58]	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2
TPL [65]	83.4	62.2	26.4	71.8	18.2	49.5	66.5	63.8	73.4	19.0	56.6	35.7	69.3	61.3	71.7	69.2	39.1	66.3	44.8	35.9	45.5	53.8
MCOF [42]	86.8	73.4	26.6	60.6	31.8	56.3	76.0	68.9	79.4	18.8	62.0	36.9	74.5	66.9	74.9	58.1	44.6	68.3	36.2	64.2	44.0	57.6
GAIN [28]	88.2	79.3	33.7	67.9	50.5	62.5	76.0	72.2	77.6	20.3	65.8	19.5	72.6	73.0	75.2	71.4	42.4	72.8	21.4	61.5	48.6	59.6
AffinityNet [20]	88.0	61.1	29.2	73.0	40.5	54.1	75.2	70.4	75.1	27.8	62.5	51.4	78.4	68.3	76.2	71.8	40.7	74.9	49.2	55.0	48.3	60.5
MDC [22]	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
OAA	90.1	77.8	36.0	80.6	49.9	61.4	73.6	73.5	78.5	21.4	68.6	28.2	73.3	72.3	79.0	73.4	44.3	74.5	27.7	63.5	53.9	62.0
OAA$^+$	90.3	77.0	35.4	80.5	50.0	61.3	77.2	75.4	79.6	21.7	71.8	29.5	75.4	73.4	78.6	74.3	44.8	76.5	27.9	65.4	52.6	62.8
OAA++	90.3	80.6	35.2	78.8	49.9	59.9	76.4	76.6	80.3	21.1	69.2	27.6	75.5	72.6	78.8	74.5	46.6	75.8	28.1	67.2	52.6	62.7
OAA++$^+$	90.3	81.5	36.8	76.7	48.9	61.1	78.7	75.1	80.2	20.6	70.7	27.6	76.4	75.6	79.7	75.1	45.6	76.7	28.5	68.7	52.8	63.2
Upperbound	92.7	86.3	37.4	79.8	61.8	68.5	87.7	81.3	84.7	30.3	76.9	61.5	80.0	75.2	81.9	80.6	55.4	81.6	53.6	76.1	62.4	71.2

online attention accumulation strategy, which increases the range of attention movement to accumulate more object regions. Plugging the attention drop layer into our framework further improves OAA and OAA$^+$ a lot on both the validation and test sets. Additionally, we also show the segmentation results based on ResNet-101 [5] backbone. Our proposed approach achieves the best result on the PASCAL VOC 2012 test set. We also provide detailed IoU scores for each category in Tab. 2 and Tab. 3.

4.3 Ablation Analysis

This section performs a series of ablation experiments and gives a detailed analysis to demonstrate the proposed strategies' effectiveness. All the segmentation results are evaluated on a single scale. Note that we use the VGGNet based DeepLab-LargeFOV model in this subsection. Furthermore, we demonstrate how the produced attention maps can benefit the semantic segmentation task. We also present some details about the evolution of attention maps during different training times.

Accumulation Strategies. The attention fusion strategy is used in OAA to accumulate the discovered discriminative regions in 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. 4, using attentions by CAM [25] 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 [25]. Moreover, OAA with the maximum fusion strategy is more effective than that with the average fusion strategy. This result is because the averaging 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.

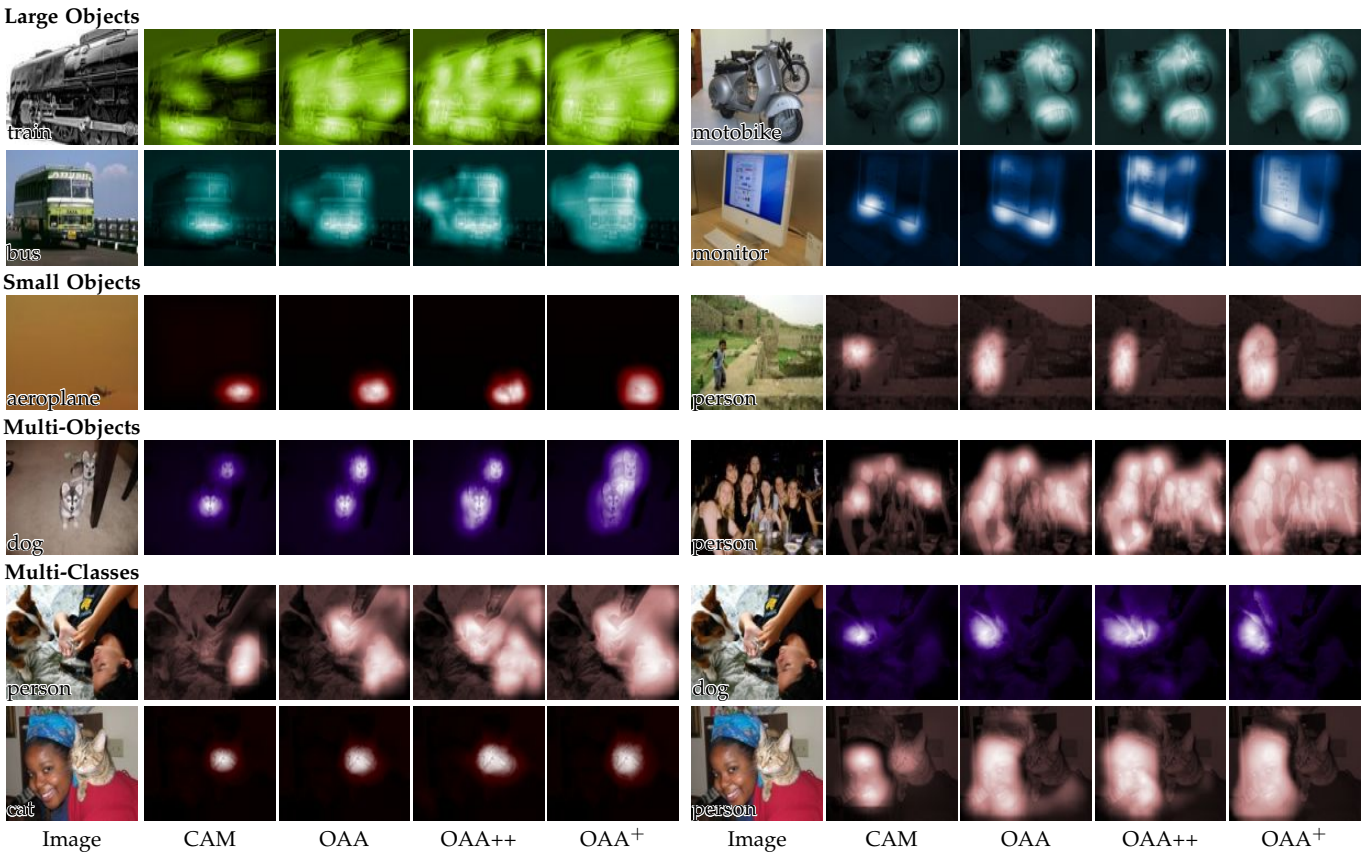


Fig. 7. Visual comparisons among different attention maps produced by CAM [25], OAA, OAA++, and OAA$^+$. OAA++ denotes OAA with the attention drop layer. OAA$^+$ denotes the integral attention model. Best viewed in color.

Table 4
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 Eqn. (6). **EP:** OAA$^+$ only using the proposed enhanced loss in Eqn. (7). **HL:** OAA$^+$ using the proposed hybrid loss in Eqn. (9).

No.	AVE	MAX	MCE	EP	HL	mIoU (%)
1						53.9
2	✓					57.0
3		✓				58.6
4		✓	✓			51.2
5		✓		✓		53.4
6		✓			✓	59.6

Table 5
Segmentation results using the hybrid loss with different thresholds for N_+^c. We compute the noise rate and recall of the attention maps generated from OAA$^+$. The attention map is thresholded to a binary mask with a threshold of 50% of the maximum attention value.

No.	Threshold	Noise (%)	Recall (%)	mIoU (%)
1	0.0	43.9	59.4	59.6
2	0.1	40.1	54.9	59.5
3	0.2	37.4	51.0	58.9
4	0.3	35.4	47.4	58.4

Loss Function in OAA$^+$. As stated in Sec. 3.4, 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. 4, 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) [58] 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. We also conduct an ablation experiment about the hybrid loss, where we test the result of only using the proposed enhanced loss, *i.e.*, Eqn. (7). The result of using the hybrid loss outperforms that of only using the enhanced loss by 6.2%. This fact indicates that the importance of the constraints loss, *i.e.*, Eqn. (8).

We also conduct experiments to study the impact of the background noise on the hybrid loss. We utilize different thresholds to divide N_+^c and N_-^c, where the lower the threshold, the higher the background noise in N_+^c is. In Tab. 5, we can see that when decreasing the threshold from 0.3 to 0.0, the hybrid loss will introduce more background noise to the attention maps (35.4% to 43.9%). Thus, the hybrid loss is sensitive to the background noise. However, decreasing the threshold will help the attention maps mine many potential object pixels (47.4% to 59.4%). When generating the pseudo segmentation labels, we use saliency maps to generate background locations, which can help filter out some background noise. The pseudo segmentation

Table 6
Segmentation results using the hybrid loss function with different training iterations for the segmentation network.

Iterations	15000	20000	25000	30000
mIoU (%)	59.6	59.5	59.7	59.6

Table 7
Comparisons of mIoU scores on PASCAL VOC 2012 validation set when using a 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.

No.	Training Images	Proportion	mIoU (%)
1	2,116 weak	20	54.6
2	5,291 weak	50	57.3
3	8,466 weak	80	58.9
4	10,582 weak	100	59.6
5	8,466 weak + 2,116 pixel	-	61.6
6	5,291 weak + 5,291 pixel	-	63.7
7	2,116 weak + 8,466 pixel	-	65.1
8	10,582 pixel	-	66.1

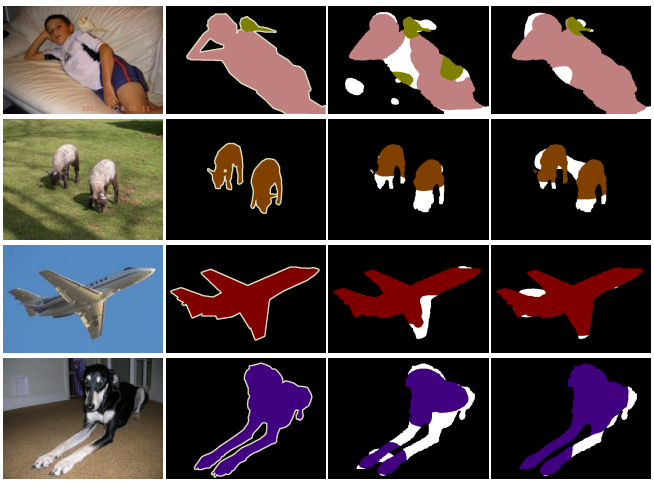


Fig. 8. Comparison of pseudo segmentation labels. (a) Source images; (b) Ground-truth segmentation labels; (c) Pseudo segmentation labels based on attention maps generated from the final model trained with the attention drop layer; (d) Pseudo segmentation labels based on attention maps generated from our OAA with attention drop layer.

Table 8
Ablation on the hyper-parameters of the attention drop layer.

No.	δ	β	mIoU (%)
1	0.1	0.5	59.2
2	0.4	0.5	59.4
3	0.5	0.5	59.7
4	0.6	0.5	**59.9**
5	0.7	0.5	59.4
6	0.8	0.5	58.9
7	0.6	0.3	59.4
8	0.6	0.1	59.3
9	0.6	0.7	59.3
10	0.6	0.9	59.1

labels are not vulnerable to background noise but are more susceptible to the recall of the target objects. Thus, the segmentation result increases a lot when decreasing the threshold from 0.3 to 0.0.

Additionally, we conduct experiments to study whether the hybrid loss is sensitive to the training steps. We train the segmentation network with different training iterations. As shown in Tab. 6, the performance slightly changes ($\pm 0.1\%$) for different training iterations, which verifies that the hybrid loss is not sensitive to the training steps of the segmentation network.

Results with Different Strategies. As shown in Tab. 4, 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. 4, it can be seen that using OAA$^+$ can further improve the results by OAA by 1.0% on the validation set. This result indicates our integral attention model with the proposed hybrid loss function can improve the quality of the cumulative attention maps.

The Number of Training Images. To further investigate the quality of the proxy segmentation labels using our attention maps, we attempt to use different numbers of segmentation labels to train the segmentation network. We use the attention maps produced by OAA$^+$ to generate the proxy segmentation labels.

As shown in Tab. 7, 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 [25]. This result indirectly suggests that our attention maps are of high quality and facilitate the segmentation task. We also use both the proxy segmentation labels and ground-truth labels to train the segmentation models. With more ground-truth labels are used for training, the mIoU scores increase, which demonstrates the proxy segmentation labels still have a large room for improvement.

Visual Comparisons. As shown in Fig. 7, we show some qualitative results and give the corresponding attention maps produced by CAM [25], OAA, OAA++, and OAA$^+$, respectively, for visual comparisons. The images shown include different 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 than the attention maps produced by CAM [25]. On the fourth 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. In the last two rows, we show some examples containing multiple classes. Our cumulative attention maps can successfully distinguish different classes and detect the target objects densely. The attention maps produced by OAA++ and OAA$^+$ can discover more integral object regions than the cumulative attention maps from OAA. Additionally, we also show some segmentation results in Fig. 9.

Hyper-parameters of Attention Drop Layer. When integrating the attention drop layer, there are two hyperparameters: the attention threshold (δ) and drop rate (β). We conduct ablation experiments to analyze their effects on the

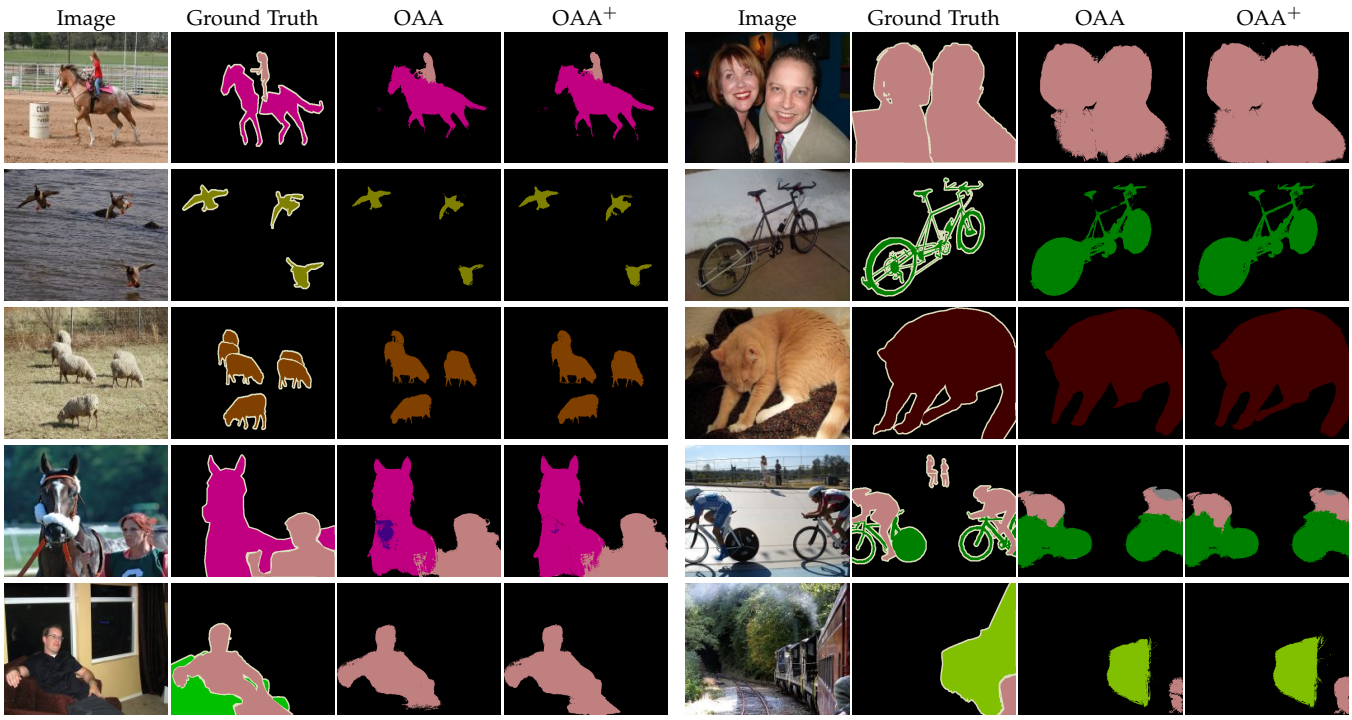


Fig. 9. Qualitative segmentation results on the PASCAL VOC 2012 validation set using attention maps generated by our OAA and OAA$^+$, respectively. We also show several failure cases on the bottom row.

Table 9
Comparisons of the variants of OAA with the attention drop layer on PASCAL VOC 2012 validation sets. **OAA-drop**: OAA with the attention drop layer; **OAA-drop-feat**: OAA with the attention drop layer on features; **OAA-drop-model**: attention maps from the model trained with the attention drop layer. * denotes the results are from [19].

Strategy	mIoU (%)
OAA	58.6
OAA-drop	59.9
OAA-drop-feat	59.5
OAA-drop-model	58.3
SeeNet* [19]	57.3
ACoL* [27]	56.1

Table 10
Segmentation results using the attention drop layer with different training iterations.

Iterations	20000	30000	45000	60000
mIoU (%)	59.4	59.9	59.7	59.8

segmentation results. We fix the value of one parameter when we study another parameter. When studying δ, β is set to 0.5. When studying β, we set δ to 0.6 because it achieves the best performance.

As shown in Tab. 8, when δ approaches 1, the improvement brought by the drop layer decreases largely (59.9% *v.s.* 58.9%), When δ is set to a large value, the drop region becomes very small, causing the mIoU scores are close to OAA without the drop layer. When δ comes to small values, we can observe the segmentation results also decrease. This result is because the attention regions may contain large object regions when δ is set to a small value. The drop layer may force attention shift to non-target regions, causing attention with much noise. For the drop rate β, when β is set to 0.5, the segmentation result achieves the best performance.

The Variant of OAA-drop. In Sec. 3.3, we propose integrating an attention drop layer into OAA (OAA++, denote as OAA-drop here), which can further improve the cumulative attention map's quality. The drop layer occludes the regions with strong attention values in input images. We also test the performance when occluding the regions in features (OAA-drop-feat). We select the features from the last convolutional layer in the classification network. As shown in Tab. 9, occluding features' performance is slightly lower than occluding the input images (59.9% *v.s.* 59.5%).

Additionally, our attention drop layer is based on the erasing-based strategy. So we compare our method to the erasing-based methods [19], [27], which all generate attention maps from the final classification models. Compared with the erasing-based methods, such as SeeNet [19], OAA outperforms them by a large margin. We also generate the attention maps from the final classification model trained with the attention drop layer (OAA-drop-model). The segmentation result using cumulative attention maps from OAA-drop outperforms that using attention maps from OAA-drop-model by 1.6%. In Fig. 8, we also present several pseudo segmentation labels generated using the attention maps of OAA-drop and OAA-drop-model, respectively. The pseudo segmentation labels from our OAA-drop are more accurate and complete. These facts all verify the effectiveness of the online attention accumulation process and the attention drop layer.

Training of Attention Drop Layer. We have explored the impact of different training iterations of the classification network on segmentation performance when integrating

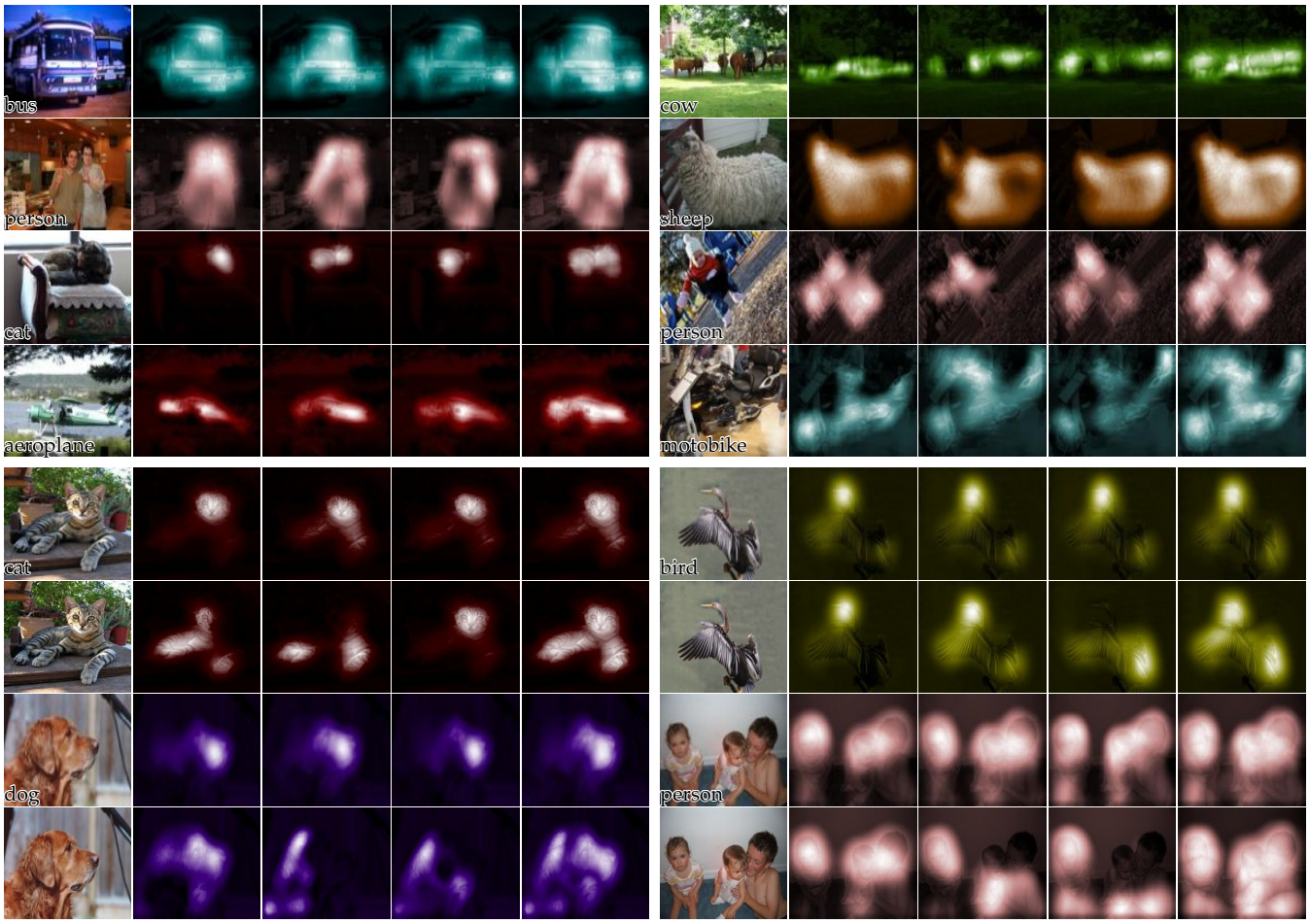


Fig. 10. The evolution of attention maps during training. (a) Images; (b-d) Intermediate attention maps; (e) Cumulative attention maps. The attention maps at the first four rows are generated from the normal training process. The attention maps at the last four rows are generated from the normal training process and training process with the attention drop layer, respectively.

Table 11
Comparisons of the attention maps' quality from different strategies.

Strategy	CAM	OAA	OAA-drop
mIoU (%)	53.9	58.6	59.9
$recall$ (%)	36.7	52.3	61.9
$1-precision$(%)	32.7	37.4	41.1

Table 12
Segmentation results based on different segmentation networks. The segmentation results are evaluated with the multi-scale test and CRF post process.

Methods	DeepLab-LargeFOV [70]	DeepLabv2 [7]
OAA	63.9	67.6
OAA++	64.9	68.4
OAA$^+$	65.2	68.2
OAA++$^+$	66.1	68.9
UpperBound	75.4	77.9

the attention drop layer into OAA. In Tab. 10, we can see that the performance is improved from 59.4% to 59.9% when increasing the training iterations from 20000 to 30000. We analyze that when the training iterations are set to 20000, the classification model doesn't converge. Thus, when increasing the training iterations from 20000 to 30000, attention still shifts on the target objects. When further increasing the training iterations to 45000/60000, the performance is not improved. The extra training iterations cannot mine new object regions. We analyze that this is because attention doesn't shift when the classification model converges.

Quality of Cumulative Attention Maps. We utilize the $recall$ and $precision$ metric to study the quality of cumulative attention maps. Specifically, the attention map is first thresholded to a binary mask with a threshold of 50% of the maximum attention value. Compared with the segmentation annotations, we compute $recall$ to denote the ratio of discovered target regions and $1-precision$ to denote the ratio of the noisy regions. It can be seen that OAA achieves a higher recall without introducing much noise. When the attention drop layer is integrated into OAA, the recall can be further improved. This comparison demonstrates the quality of the cumulative attention maps.

More Advanced Segmentation Network DeepLabv2 [70] is an improved version of DeepLab-LargeFOV [7]. DeepLabv2 employs the atrous spatial pyramid pooling module to segment objects at multiple scales. We have performed the segmentation experiments using DeepLabv2. As shown in Tab. 12, when using DeepLabv2 instead of DeepLab-LargeFOV, the performance is largely improved.

Table 13
Comparison of the recall of the target object regions discovered by attention maps.

Methods	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	avg
CAM [25]	55.2	43.0	45.4	46.3	43.4	32.5	38.1	22.7	31.9	39.5	31.7	30.0	34.9	37.7	35.9	42.4	45.2	26.7	31.2	46.9	36.7
OAA	68.2	60.4	58.0	59.5	57.9	49.2	59.2	34.2	51.4	51.4	47.6	39.2	52.7	54.0	54.3	55.7	61.7	42.5	38.4	60.4	52.3
OAA++	78.5	69.2	73.1	67.3	68.0	59.6	69.0	49.9	58.4	60.1	57.1	51.4	62.7	61.6	62.9	64.8	70.1	50.0	47.7	65.0	61.9

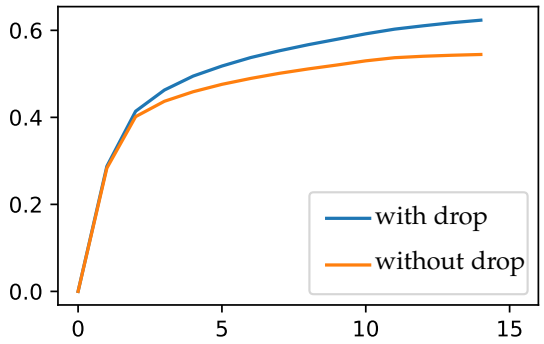


Fig. 11. Evolution of the attention maps during different training stages. The curve denotes the ratio of the founded object regions to entire object regions.

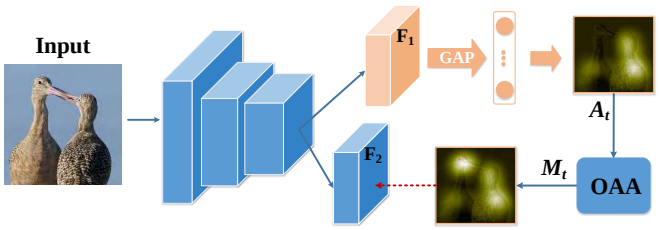


Fig. 12. An illustration of our OAA for the weakly supervised object localization. The red dotted line denotes the supervision from the cumulative attention maps of OAA.

Table 14
Comparison of the localization accuracy among different methods on CUB-200-2011 test set. These methods all generate attention maps from the VGG-16 classification network [4].

Methods	top-1 acc. (%)	top-5 acc. (%)
CAM [25]	41.46	51.36
ACoL [27]	45.92	56.51
SPG [41]	46.64	57.72
ADL [31]	52.36	-
Ours	**56.23**	**69.68**

4.4 Attention Evolution during Training

The attention constantly shifts over the target object regions during different training stages. We perform an ablation experiment to study the evolution of attention maps as the training goes on. We compute the target object regions' ratio found by the cumulative attention maps to the entire target object regions at different epochs. The target object regions are extracted from the segmentation labels.

As shown in Fig. 11, we present the evolution process of the founded object regions' ratio to the entire target object regions during training. The discovered target object regions gradually become large during training regardless of OAA with or without the attention drop layer. It can be seen that the newly discovered object regions decrease during the late epochs of the training phase. We analyze that when the image classifier tends to be convergent, its parameters fluctuate in a small range, causing the attention does not change a lot on the object regions. When integrating the attention drop layer into the OAA framework, the cumulative attention maps can discover more object regions as the training process goes on. Occluding part of discriminative regions can efficiently force attention shifting to new object regions, expanding the range of attention movement. Tab. 13 shows the recall of the attention maps from CAM, OAA, and OAA++ of each class.

In Fig. 10, we present more visual examples of attention evolution during training. At the first four rows of Fig. 10, we can see the attention maps during training, *i.e.*, Fig. 10(b-d), focus on different object parts and the final cumulative attention maps, *i.e.*, Fig. 10(e), can find more integral object regions. At the last four rows, we also present that the attention maps of some images focus on similar regions during training, causing the final cumulative attention maps to only obtain small object regions. With the attention drop layer's help, the range of attention movement becomes larger, facilitating the cumulative attention maps mine integral object regions.

4.5 Weakly Supervised Object Localization

The goal of the weakly supervised object localization is to generate a tight bounding box for the target object. Most weakly supervised object localization approaches [25], [27], [31], [41] utilize attention maps to locate the target objects. They usually threshold the attention maps to binary maps and generate bounding boxes for the target objects. As our OAA is used for the training process, we slightly adjust our method. We combine OAA and the hybrid loss into a single network to generate attention maps for test images. Specifically, we utilize two branches in a single network, as shown in Fig. 12. One branch ($F1$) is used to accumulate attention maps during training. Another branch ($F2$) is supervised by the supervision from the cumulative attention maps. At the inference time, we utilize the $F2$ branch to generate attention maps for test images.

We use the same network architecture, *i.e.*, VGG-16, as the weakly supervised semantic segmentation. The localization experiments are performed on CUB-200-2011 dataset [71] with 200 categories, which have 5,994 training images and 5,794 test images. We follow the way in [25] to generate predicted bounding boxes. For evaluating the localization performance, we utilize the Top-1 localization accuracy and Top-5 localization accuracy. As shown in Tab. 14, our OAA outperforms the baseline, CAM [25], by a large margin. Compared to the state-of-the-art localization methods, our OAA achieves superior localization accuracy, demonstrating the effectiveness of OAA for object localiza-

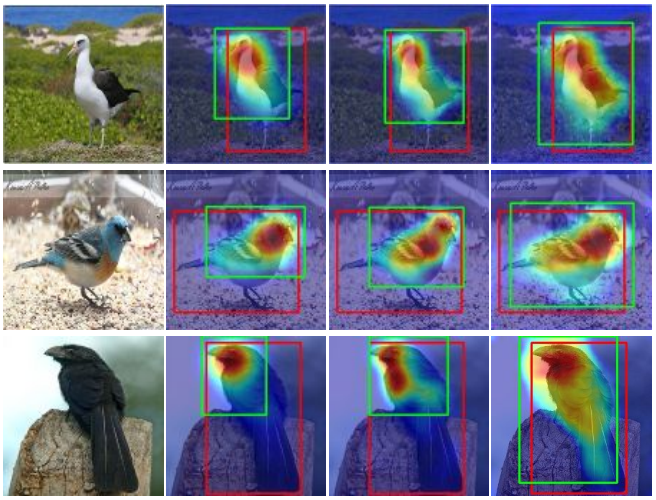


Fig. 13. Qualitative localization results on the CUB-200-2011 dataset. The ground-truth bounding box is in red. The predicted bounding box is in green.

tion. Qualitative results are shown in Fig. 13.

5 Conclusion

This paper explores 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. We also rethink the OAA strategy and propose integrating an attention drop layer into OAA, expanding the range of attention movement. Our approach is easy to follow and can be plugged into any classification network 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-art methods. Moreover, we also analyze the evolution of attention during training.

In the future, we plan to extend our method to deal with much larger datasets, *e.g.* ImageNet [72]. By fully exploring category-agnostic common features as prior knowledge, *e.g.* edges [73], salient objects [44], [74], super-pixels [75], and object proposals [76], there are potentials to further release the requirement of human labeling, by automatically learning from Internet retrieved images [77].

Acknowledgment

This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, NSFC (61922046), S&T innovation project from Chinese Ministry of Education, and the Fundamental Research Funds for the Central Universities (Nankai University, NO. 63213090).

References

[1] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong, "Integral object mining via online attention accumulation," in *Int. Conf. Comput. Vis.*, 2019, pp. 2070–2079.

[2] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inform. Process. Syst.*, 2019.

[3] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, "Jittor: a novel deep learning framework with meta-operators and unified graph execution," *Science China Information Sciences*, vol. 63, no. 12, pp. 1–21, 2020.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

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

[6] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, 2021.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *Int. Conf. Learn. Represent.*, 2015.

[8] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1925–1934.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2881–2890.

[11] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7151–7160.

[12] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4003–4012.

[13] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[14] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in *Eur. Conf. Comput. Vis.*, 2016, pp. 90–105.

[15] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 3159–3167.

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

[17] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1713–1721.

[18] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari, "Training object class detectors from eye tracking data," in *Eur. Conf. Comput. Vis.*, 2014, pp. 361–376.

[19] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 549–559.

[20] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4981–4990.

[21] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7014–7023.

[22] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7268–7277.

[23] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Int. Conf. Comput. Vis.*, 2015, pp. 1796–1804.

[24] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Eur. Conf. Comput. Vis.*, 2016, pp. 695–711.

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2921–2929.

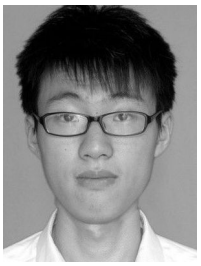
[26] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classifi-

cation to semantic segmentation approach," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1568–1576.
[27] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1325–1334.
[28] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Guided attention inference network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 2996–3010, 2019.
[29] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *Int. Conf. Mach. Learn.*, 2017, pp. 233–242.
[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[31] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2219–2228.
[32] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.
[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization." in *Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[34] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.
[35] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2020, pp. 24–25.
[36] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Int. Conf. Comput. Vis.*, 2017, pp. 3429–3437.
[37] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Int. Conf. Comput. Vis.*, 2019, pp. 2950–2958.
[38] S.-A. Rebuffi, R. Fong, X. Ji, and A. Vedaldi, "There and back again: Revisiting backpropagation saliency methods," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8839–8848.
[39] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *Eur. Conf. Comput. Vis.*, 2016, pp. 543–559.
[40] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 9215–9223.
[41] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Eur. Conf. Comput. Vis.*, 2018, pp. 597–613.
[42] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1354–1362.
[43] Q. Hou, D. Massiceti, P. K. Dokania, Y. Wei, M.-M. Cheng, and P. H. Torr, "Bottom-up top-down cues for weakly-supervised semantic segmentation," in *Int. Worksh. on Energy Minimization Methods in Comput. Vis. Pattern Recog.*, 2017, pp. 263–277.
[44] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, 2019.
[45] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, 2017.
[46] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.
[47] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3917–3926.
[48] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *Int. Conf. Learn. Represent.*, 2015.
[49] X. Wang, L. Zhu, Y. Wu, and Y. Yang, "Symbiotic attention for egocentric action recognition with object-centric alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
[50] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic attention with privileged information for egocentric action recognition," in *AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 12 249–12 256.
[51] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *Int. Conf. Learn. Represent.*, 2015.
[52] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 642–651.
[53] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Int. Conf. Comput. Vis.*, 2015, pp. 1742–1750.
[54] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 12 275–12 284.
[55] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Weakly-supervised semantic segmentation via subcategory exploration," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8991–9000.
[56] G. Sun, W. Wang, J. Dai, and L. Van Gool, "Mining cross-image semantics for weakly supervised semantic segmentation," in *Eur. Conf. Comput. Vis.*, 2020, pp. 347–365.
[57] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Int. Conf. Comput. Vis.*, 2019, pp. 5208–5217.
[58] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, 2017.
[59] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in *Eur. Conf. Comput. Vis.*, 2016, pp. 218–234.
[60] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3529–3538.
[61] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5038–5047.
[62] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 7322–7330.
[63] B. Jin, M. V. Ortiz Segovia, and S. Susstrunk, "Webly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3626–3635.
[64] A. Chaudhry, P. K. Dokania, and P. H. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in *Brit. Mach. Vis. Conf.*, 2017, pp. 20.1–20.13.
[65] D. Kim, D. Yoo, I. S. Kweon *et al.*, "Two-phase learning for weakly supervised object localization," in *Int. Conf. Comput. Vis.*, 2017, pp. 3534–3543.
[66] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," in *Eur. Conf. Comput. Vis.*, 2018, pp. 367–383.
[67] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Adv. Neural Inform. Process. Syst.*, vol. 24, 2011, pp. 109–117.
[68] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Int. Conf. Comput. Vis.*, 2011, pp. 991–998.
[69] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[70] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.
[71] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.
[72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
[73] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939 – 1946, 2019.

[74] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, no. 2, pp. 117–150, 2019.
[75] J. Zhao, R. Bo, Q. Hou, M.-M. Cheng, and P. Rosin, "Flic: Fast linear iterative clustering with active search," *Computational Visual Media*, vol. 4, no. 4, pp. 333–348, Dec 2018. [Online]. Available: https://doi.org/10.1007/s41095-018-0123-y
[76] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. S. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," *Computational Visual Media*, vol. 5, no. 1, pp. 3–20, Mar 2019. [Online]. Available: https://doi.org/10.1007/s41095-018-0120-1
[77] Q. Hou, L. Han, and M.-M. Cheng, "Autonomous learning of semantic segmentation from internet images (in chinese)," *Sci Sin Inform*, 2021.

Peng-Tao Jiang is a Ph.D. student from the College of Computer Science at Nankai University, under Prof. Ming-Ming Cheng's supervision. Before that, he received a Bachelor Degree from Xidian University in 2017. His research interests include weakly supervised tasks and model interpretability.

Ling-Hao Han is currently an master student at the College of Computer Science, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning and computer vision.



Qibin Hou received his Ph.D. degree from the School of Computer Science, Nankai University, under Prof. Ming-Ming Cheng's supervision. Currently, he is a research fellow working with Prof. Jiashi Feng at the National University of Singapore. His research interests include deep learning, image processing, and computer vision.

Ming-Ming Cheng is a professor at Nankai University, leading the Media Computing Lab. He received his Ph.D. degree from Tsinghua University in 2012. Then he did two years research fellow with Prof. Philip Torr in Oxford. Prof. Cheng's research interests include computer vision and image processing. He received research awards, including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program. He is on the editorial boards of IEEE TIP.

Yunchao Wei is currently an Assistant Professor at the University of Technology Sydney. He received his Ph.D. degree from Beijing Jiaotong University in 2016. Before joining UTS, he was a Postdoc Researcher in Prof. Thomas Huang's Image Formation and Professing (IFP) group at Beckman Institute, UIUC, from 2017 to 2019. His research interests mainly include Deep learning and its applications in computer vision, e.g., image classification, learning with imperfect data.