

Leveraging Auxiliary Tasks with Affinity Learning for Weakly Supervised Semantic Segmentation

Lian Xu¹, Wanli Ouyang², Mohammed Bennamoun¹, Farid Boussaid¹, Ferdous Sohel³, and Dan Xu⁴

¹The University of Western Australia ²The University of Sydney

³Murdoch University ⁴Hong Kong University of Sciences and Technology

{lian.xu,mohammed.bennamoun,farid.boussaid}@uwa.edu.au, wanli.ouyang@sydney.edu.au,
F.Sohel@murdoch.edu.au, danxu@cse.ust.hk

Abstract

Semantic segmentation is a challenging task in the absence of densely labelled data. Only relying on class activation maps (CAM) with image-level labels provides deficient segmentation supervision. Prior works thus consider pre-trained models to produce coarse saliency maps to guide the generation of pseudo segmentation labels. However, the commonly used off-line heuristic generation process cannot fully exploit the benefits of these coarse saliency maps. Motivated by the significant inter-task correlation, we propose a novel weakly supervised multi-task framework termed as **AuxSegNet**, to leverage saliency detection and multi-label image classification as auxiliary tasks to improve the primary task of semantic segmentation using only image-level ground-truth labels. Inspired by their similar structured semantics, we also propose to learn a cross-task global pixel-level affinity map from the saliency and segmentation representations. The learned cross-task affinity can be used to refine saliency predictions and propagate CAM maps to provide improved pseudo labels for both tasks. The mutual boost between pseudo label updating and cross-task affinity learning enables iterative improvements on segmentation performance. Extensive experiments demonstrate the effectiveness of the proposed auxiliary learning network structure and the cross-task affinity learning method. The proposed approach achieves state-of-the-art weakly supervised segmentation performance on the challenging PASCAL VOC 2012 and MS COCO benchmarks.¹

1. Introduction

Semantic segmentation plays a vital role in many applications such as scene understanding and autonomous driv-

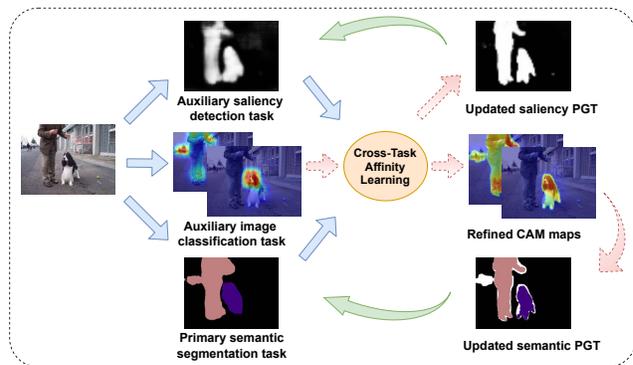


Figure 1. An illustration of the proposed approach for weakly supervised semantic segmentation. Our approach jointly learns two auxiliary tasks (*i.e.*, multi-label image classification and saliency detection) and a primary task (*i.e.*, semantic segmentation) only using image-level ground-truth labels, and performs affinity learning across two dense prediction tasks (*i.e.*, saliency detection and semantic segmentation). The learned affinity is then used to generate updated pseudo ground-truth (PGT) providing supervision to learn saliency detection and semantic segmentation.

ing. It describes the process of assigning a semantic label to each pixel of an image. Prior works have achieved great success in the case of fully supervised semantic segmentation using Convolutional Neural Networks (CNNs). However, this has come at a high pixel-wise annotation cost. There has been an emerging research trend in semantic segmentation using less expensive annotations, such as bounding boxes [14, 31], scribbles [22, 33], points and image-level labels [28, 18, 42]. Among them, image-level labels only indicate the presence or absence of objects in an image, resulting in an inferior segmentation performance compared to their fully supervised counterparts.

Most existing weakly supervised semantic segmentation (WSSS) approaches follow a two-step pipeline, *i.e.*, generating pseudo segmentation labels and training segmenta-

¹<https://github.com/xulianuwa/AuxSegNet>

tion models. A key element in generating pseudo segmentation labels is the class activation map (CAM) [48], which is extracted from CNNs trained on image-level labels. Although CAM maps identify class-specific discriminative image regions, those regions are quite sparse with very coarse boundaries. In order to generate high-quality pseudo segmentation labels, many approaches [37, 38, 17, 36] have been presented to improve CAM maps from various aspects. Besides, existing methods [4, 13, 32, 47] typically use off-the-shelf saliency maps, combined with CAM maps, to determine reliable object and background regions. A general pre-trained saliency model can provide coarse saliency maps, which contain useful object localization information, on a target dataset. However, in prior works, such coarse off-the-shelf saliency maps are only used as fixed binary cues in an off-line pseudo label generation process via heuristic thresholding, they are neither directly involved in the network training nor updated, which largely restricts their use to further benefit the segmentation performance.

Motivated by the observation that semantic segmentation, saliency detection and image classification are highly correlated, we propose a weakly supervised multi-task deep network (see Figure 1), which leverages saliency detection and multi-label image classification as auxiliary tasks to help learn the primary task of semantic segmentation. Through the joint training of these three tasks, an online adaptation can be achieved from pre-trained saliency maps to our target dataset. In addition, the task of saliency detection impels the shared knowledge to emphasize the difference between foreground and background pixels, thus driving the object boundaries of the segmentation outputs to coincide with those of the saliency outputs. Similarly, the image classification task highlights the discriminative features which can lead to more accurate segmentation predictions.

Furthermore, we notice that, similar to these two dense prediction tasks, *i.e.*, semantic segmentation and saliency detection, CAM maps also represent pixel-level semantics albeit they are only partially activated. Therefore, we propose to learn global pixel-level pair-wise affinities from the features of segmentation and saliency tasks to guide the propagation of CAM activations. More specifically, two task-specific affinity maps are first learned for the saliency and segmentation tasks, respectively. To capture the complementary information between the two affinity maps, they are then adaptively integrated based on the learned self-attentions to produce a cross-task affinity map. Moreover, as we expect to learn semantic-aware and boundary-aware affinities so as to better update pseudo labels, we impose constraints on learning the cross-task affinities from task-specific supervision and joint multi-objective optimization. The learned cross-task affinity map is further utilized to refine saliency predictions and CAM maps to provide improved pseudo labels for both saliency detection and seman-

tic segmentation respectively, enabling a multi-stage cross-task iterative learning and label updating.

In summary, the main contribution is three-fold:

- We propose an effective multi-task auxiliary deep learning framework (*i.e.*, AuxSegNet) for weakly supervised semantic segmentation. The proposed AuxSegNet leverages multi-label image classification and saliency detection as auxiliary tasks to help learn the primary task (*i.e.*, semantic segmentation) using only image-level ground-truth labels.
- We propose a novel method to learn cross-task affinities to refine both task-specific representations and predictions for semantic segmentation and saliency detection. The learned global pixel-level affinities can also be used to simultaneously update semantic and saliency pseudo labels in a joint cross-task iterative learning framework, yielding continuous boosts of the semantic segmentation performance.
- Our proposed method achieves state-of-the-art results on PASCAL VOC 2012 and MS COCO datasets for the task of weakly supervised semantic segmentation.

2. Related Work

In this section, we review recent works from two closely related perspectives, *i.e.*, weakly supervised semantic segmentation and auxiliary learning for segmentation.

Weakly supervised semantic segmentation. Recent WSSS approaches commonly rely on CAM maps as seeds to generate pseudo segmentation labels. Several works focus on modifying classification objectives [36] or network architectures [38] to improve CAM maps. A few works mine object regions based on erasing [37, 13, 20] and accumulation [17] heuristics. Moreover, improved CAM maps can be achieved by exploring sub-category information [3] or mining cross-image semantics [10, 32, 21].

There are several methods [37, 35, 34] which also perform iterative refinement on pseudo segmentation labels. Wei *et al.* [37] iteratively train the CAM network using images of which the discovered object parts are erased, to mine more object regions. This still results in combined CAM maps having coarse boundaries and non-discriminative object parts not activated. Wang *et al.* [35] improve pseudo segmentation labels by iteratively mining common object features from superpixels. Wang *et al.* [34] use an affinity network, which learns local pixel affinities, to refine pseudo segmentation labels. The affinity network is iteratively optimized using the results from a segmentation network. In contrast to these methods requiring learning an additional single-modal affinity network to have alternating training with the segmentation network, we perform

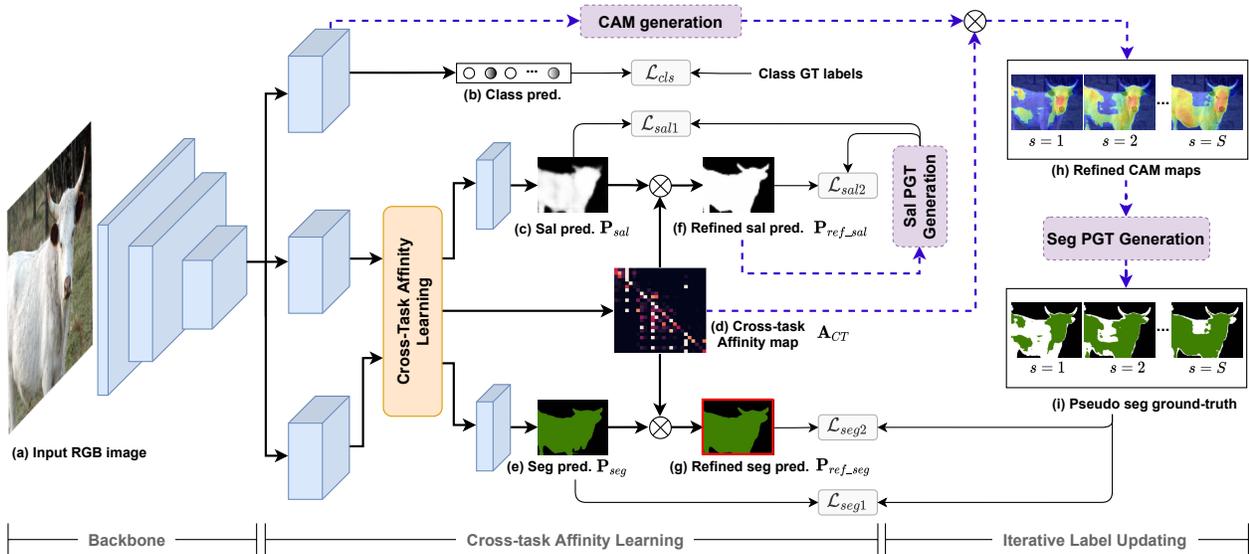


Figure 2. An overview of the proposed AuxSegNet. An input RGB image (a) is first passed through a backbone network to extract image features, which are then fed to three branches for multi-label image classification (b), saliency detection (c & f) and semantic segmentation (e & g), respectively. The proposed cross-task affinity learning module (see Figure 3) takes as inputs the segmentation and saliency feature maps, and outputs augmented feature maps for predicting both tasks (c & e) and a cross-task affinity map (d) for task-specific prediction refinement (f & g). The refined saliency predictions are used to update pseudo saliency labels, and the learned cross-task affinity map is used to refine CAM maps (h) to update pseudo segmentation labels (i) to retrain the network. The network training (black solid lines) and label updating (blue dashed lines) are performed alternatively for multiple stages (*i.e.*, $s = 1, 2, \dots, S$) to learn a more reliable affinity map and produce more accurate segmentation predictions.

the cross-task affinity learning simultaneously with the proposed joint multi-task auxiliary learning network.

Several other works [1, 10] also refine pseudo labels by learning pixel affinities. However, Ahn *et al.* [1] only learn pixel affinities based on the selected samples from the sparse CAM maps. Fan *et al.* [10] learn pixel affinity only to enhance feature representations for object estimation. In contrast, the proposed method is different from these approaches in the following aspects: it learns global pixel affinities across different tasks; it learns the pixel-level affinities to refine both task-specific representations and predictions; the affinity can be progressively improved along with more accurate saliency and segmentation results to be achieved with updated pseudo labels on both tasks.

Auxiliary learning for segmentation. Many existing works have shown the effectiveness of multi-task learning [40, 30], which allows the sharing of learned knowledge across tasks to improve the performance of each individual task. In auxiliary learning, the goal of auxiliary tasks is to improve the performance of the primary task [41, 25]. For instance, Dai *et al.* [7] propose a multi-task network for instance segmentation by jointly learning to differentiate instances, estimate masks, and categorize objects. In [5], more accurate segmentation is achieved by learning an auxiliary contour detection task. In these cases, ground-truth labels are provided for both primary and auxiliary tasks.

In weakly supervised learning, the joint learning of ob-

ject detection and segmentation has been explored in [29, 16]. The joint learning of image classification and semantic segmentation has been investigated in [4, 44, 2]. Zeng *et al.* [43] recently propose a joint learning framework for saliency detection and weakly supervised semantic segmentation, which however uses strong pixel-level saliency ground-truth labels as supervision. In contrast, we leverage two auxiliary tasks (*i.e.*, image classification and saliency detection) to facilitate the learning of the primary task of semantic segmentation using image-level ground-truth classification labels and off-the-shelf saliency maps. We take further advantage of multi-task features to learn cross-task affinities, which can improve the pseudo labels for both saliency and segmentation tasks simultaneously to achieve progressive boosts of the segmentation performance.

3. The Proposed Approach

Overview. The overall architecture of the proposed AuxSegNet is shown in Figure 2. An input RGB image is first fed into a shared backbone network. The generated backbone features are then forwarded to three task-specific branches which predict the class probabilities, a dense saliency map, and an dense semantic segmentation map, respectively. The proposed cross-task affinity learning module first learns task-specific pixel-level pair-wise affinities, which are used to enhance the features of the saliency

and segmentation tasks, respectively. Then, the two task-specific affinity maps are adaptively integrated with the learned self-attentions as a guide to produce a global cross-task affinity map. This affinity map is further used to refine both the saliency and segmentation predictions during training. After each training stage, the learned cross-task affinity map is used to update the saliency and segmentation pseudo labels by refining the saliency predictions and the CAM maps. Only image-level ground-truth labels are required to train the proposed AuxSegNet.

3.1. Multi-Task Auxiliary Learning Framework

Auxiliary supervised image classification. The input image is first passed through the multi-task backbone network to produce a feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times K}$, where K is the number of channels, H and W are the height and width of the map, respectively. A Global Average Pooling (GAP) layer is then applied on \mathbf{F} by aggregating each channel of \mathbf{F} into a feature vector. After that, a fully connected (fc) layer is performed as a classifier to produce a probability distribution of the multi-class prediction. Given the weight matrix of the fc classifier $\mathbf{W} \in \mathbb{R}^{K \times C}$, with C denoting the number of classes, the CAM map for a specific class c at a spatial location (i, j) can be formulated as $\text{CAM}_c(i, j) = \sum_k^K \mathbf{W}_k^c \mathbf{F}_k(i, j)$ where \mathbf{W}_k^c represents the weights corresponding to the class c and the feature channel k , and $\mathbf{F}_k(i, j)$ represents the activation from the k -th channel of \mathbf{F} at a spatial location (i, j) . The generated CAM maps are then normalized to be between 0 and 1 for each class c by the maximum value in the two spatial dimensions.

Auxiliary weakly supervised saliency detection with label updating. For the saliency detection branch, feature maps from the backbone network are forwarded to two consecutive convolutional layers with dilated rates of 6 and 12, respectively. The generated feature maps $\mathbf{F}_{sal.in}$ are then fed to the proposed cross-task affinity learning module to obtain refined feature maps $\mathbf{F}_{sal.out}$ and a global cross-task affinity map \mathbf{A}_{CT} . $\mathbf{F}_{sal.in}, \mathbf{F}_{sal.out} \in \mathbb{R}^{H \times W \times D}$ with D denoting the number of channels. The refined feature maps are used to predict saliency maps \mathbf{P}_{sal} by using a 1×1 convolutional layer followed by a Sigmoid layer. The predicted saliency maps are further refined by the generated cross-task affinity map \mathbf{A}_{CT} to obtain refined saliency predictions $\mathbf{P}_{ref.sal}$. Since no saliency ground-truth is provided, we take advantage of pre-trained models which provide coarse saliency maps $\mathbf{P}t_{sal}$ as initial pseudo labels. For the following training stages, we incorporate the refined saliency predictions of the previous stage (*i.e.*, stage $s - 1$) to iteratively perform saliency label updates to continually improve the saliency learning as follows:

$$\text{PGT}_{sal}^s = \begin{cases} \mathbf{P}t_{sal} & \text{if } s = 0, \\ \text{CRF}_d\left(\frac{\mathbf{P}_{ref.sal}^{s-1} + \mathbf{P}t_{sal}}{2}\right) & \text{if } s > 0, \end{cases} \quad (1)$$

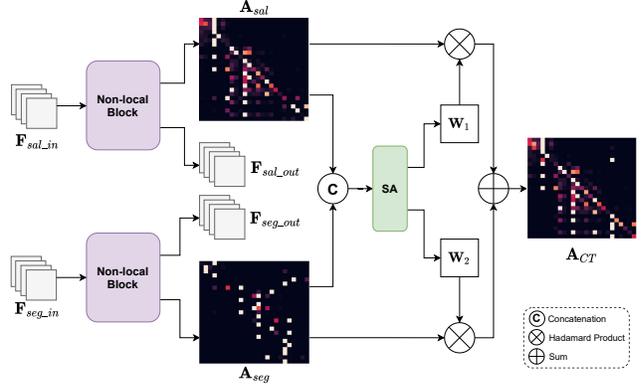


Figure 3. The detailed structure of the proposed cross-task affinity learning module. It consists of two non-local blocks which respectively learn task-specific saliency and segmentation affinity maps and refine their corresponding feature maps, and a self-attention (SA) module is designed to adaptively integrate two task-specific affinity maps to produce a global cross-task affinity map.

where PGT_{sal}^s denotes the saliency pseudo labels for the s^{th} training stage, and $\text{CRF}_d(\cdot)$ denotes a densely connected CRF following the formulation in [12] while using the average of $\mathbf{P}_{ref.sal}^{s-1}$ and $\mathbf{P}t_{sal}$ as a unary term.

Primary weakly supervised semantic segmentation with label updating. The semantic segmentation decoding branch shares the same backbone with the saliency decoding branch and the image classification branch. Similar to the saliency decoding branch, two consecutive atrous convolutional layers with rates of 6 and 12 are used to extract task-specific features $\mathbf{F}_{seg.in} \in \mathbb{R}^{H \times W \times D}$, which are then fed to the cross-task affinity learning module. The output feature maps $\mathbf{F}_{seg.out}$ are forwarded through a 1×1 convolutional layer and a Softmax layer to predict segmentation masks \mathbf{P}_{seg} , which are further refined by the learned cross-task affinity map to produce refined segmentation masks $\mathbf{P}_{ref.seg}$. To generate pseudo segmentation labels, we follow the conventional procedures [37, 38, 17, 13, 19, 32] to select reliable object regions from CAM maps and background regions from off-the-shelf saliency maps [12] by hard thresholding. More specifically, to generate the pseudo segmentation labels for the initial training stage (*i.e.*, stage 0), we first only train the classification branch using image-level labels to obtain the CAM maps. For the following training stages, we generate pseudo semantic labels by using the CAM maps refined by the cross-task affinities learned at the previous training stage.

3.2. Cross-task Affinity Learning

Given an input image, we can obtain task-specific features, *i.e.*, \mathbf{F}_{sal} and \mathbf{F}_{seg} from the saliency and segmentation branches of the proposed AuxSegNet, respectively. These feature maps for dense prediction tasks contain rich context information, and there are strong semantic relation-

ships among the feature vectors in the spatial dimensions. Such beneficial information can be used to refine CAM maps as they share highly correlated semantic structures.

Cross-task affinity. As illustrated in Figure 3, we exploit non-local self-attention blocks (NonLocal), which capture the semantic correlations of spatial positions based on the similarities between the feature vectors of any two positions, to learn task-specific pair-wise affinities for saliency detection and semantic segmentation tasks, respectively:

$$[\mathbf{F}_{sal.out}, \mathbf{A}_{sal}] = \text{NonLocal}(\mathbf{F}_{sal.in}), \quad (2)$$

$$[\mathbf{F}_{seg.out}, \mathbf{A}_{seg}] = \text{NonLocal}(\mathbf{F}_{seg.in}). \quad (3)$$

More specifically, given the saliency feature maps $\mathbf{F}_{sal.in} \in \mathbb{R}^{H \times W \times D}$, we first use three 1×1 convolutional layers to transform it into a triplet of $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, which are then flattened in the spatial dimensions to be of size $HW \times D$. To compute the pair-wise affinity, we apply the dot product between each pair of entries of \mathbf{Q} and \mathbf{K} and obtain saliency affinity matrix $\mathbf{A}_{sal} \in \mathbb{R}^{HW \times HW}$, with each row representing the similarity values of a spatial position and the rest ones. We then apply the softmax operation along each row to normalize the similarity values to be between 0 and 1. Each position of the input feature maps is then modified by attending to all positions and taking the sum of the product of all positions with their corresponding affinity values associated to the given position in the feature space. The attended feature maps with aggregated context information are finally added to the input feature maps to form the output feature maps $\mathbf{F}_{sal.out}$ with enhanced pixel-level representation. We apply the same non-local operation on the segmentation feature maps $\mathbf{F}_{seg.in}$ to generate a segmentation affinity map \mathbf{A}_{seg} and enhanced segmentation feature maps $\mathbf{F}_{seg.out}$. To learn more consistent and accurate pixel affinities, we then integrate the two task-specific affinity maps by learning a self-attention (SA) module, which consists of two convolutional layers and a softmax layer. The generated two spatial attention maps from the SA module act as two weight maps, which are used to aggregate the segmentation and saliency affinity maps into one cross-task global affinity map $\mathbf{A}_{CT} \in \mathbb{R}^{HW \times HW}$ as follows:

$$[\mathbf{W}_1, \mathbf{W}_2] = \text{SA}(\text{CONCAT}(\mathbf{A}_{sal}, \mathbf{A}_{seg})), \quad (4)$$

$$\mathbf{A}_{CT} = \mathbf{W}_1 * \mathbf{A}_{sal} + \mathbf{W}_2 * \mathbf{A}_{seg}, \quad (5)$$

where $\text{CONCAT}(\cdot)$ denotes the concatenation operation, and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{HW \times HW}$ denote the learned two spatial self-attention maps as weight maps for the saliency and segmentation affinity maps, respectively.

Multi-task constraints on cross-task affinity. To enhance the affinity learning, we consider imposing task-specific constraints on the generated cross-task affinity map. To this end, the generated cross-task affinity matrix is transposed and then enforced to refine both saliency and segmentation

predictions during training as:

$$\mathbf{P}_{ref.sal}(i, j) = \sum_k^H \sum_l^W \mathbf{A}_{CT}(i, j, k, l) \cdot \mathbf{P}_{sal}(k, l), \quad (6)$$

$$\mathbf{P}_{ref.seg}(i, j) = \sum_k^H \sum_l^W \mathbf{A}_{CT}(i, j, k, l) \cdot \mathbf{P}_{seg}(k, l). \quad (7)$$

Then the learning of the cross-task affinity can gain effective supervision from both saliency and segmentation pseudo labels. Therefore, the improvements on the updated pseudo labels can boost the affinity learning.

3.3. Training and Inference

Overall optimization objective. The overall learning objective function of AuxSegNet is the sum of the losses for the three tasks:

$$\begin{aligned} \mathcal{L}_{AuxSegNet} &= \lambda_1 \cdot \mathcal{L}_{cls} + \lambda_2 \cdot \mathcal{L}_{sal} + \lambda_3 \cdot \mathcal{L}_{seg}, \\ \mathcal{L}_{sal} &= \mathcal{L}_{sal1} + \mathcal{L}_{sal2}, \\ \mathcal{L}_{seg} &= \mathcal{L}_{seg1} + \mathcal{L}_{seg2}, \end{aligned} \quad (8)$$

where \mathcal{L}_{cls} is a multi-label soft margin loss computed between the predicted class probabilities and the image-level ground-truth labels to optimize the image classification network branch; \mathcal{L}_{sal1} and \mathcal{L}_{sal2} are binary cross entropy losses computed between the predicted/refined saliency maps and the pseudo saliency label maps to optimize the saliency detection network branch and the affinity fusion module; \mathcal{L}_{seg1} and \mathcal{L}_{seg2} are pixel-wise cross entropy losses calculated between the predicted/refined segmentation masks and the pseudo segmentation label maps, and these losses optimize the segmentation branch and the affinity fusion module; and λ_1, λ_2 and λ_3 are the loss weights for corresponding tasks.

Stage-wise training. We use a stage-wise training strategy for the entire multi-task network optimization. First, we only train the image classification branch with image-level labels for 15 epochs. The learned network parameters are then used as initialization to train the entire proposed AuxSegNet. We continually train the network for multiple stages with each training stage consisting of 10 epochs and update pseudo labels for saliency and segmentation branches after each training stage.

Inference. For inference, we use the segmentation prediction refined by the learned cross-task affinity map as the final segmentation results.

4. Experiments

4.1. Experimental Settings

Datasets. To evaluate the proposed method, we conducted experiments on PASCAL VOC 2012 [8] and MS COCO datasets [23]. PASCAL VOC has 21 classes (including a

Table 1. Performance comparison of WSSS methods in terms of mIoU(%) on the PASCAL VOC 2012 *val* and *test* sets. *: without post-processing. Sup.: supervision. I: image-level ground-truth. S: off-the-shelf saliency maps. S': saliency ground-truth.

Method	Backbone	Sup.	Val	Test
DSRG (CVPR18) [15]	ResNet101	I+S	61.4	63.2
MCOF (CVPR18) [35]	ResNet101	I+S	60.3	61.2
AffinityNet (CVPR18) [1]	ResNet38	I	61.7	63.7
SeeNet (NeurIPS18) [13]	ResNet101	I+S	63.1	62.8
FickleNet (CVPR19) [19]	ResNet101	I+S	64.9	65.3
OAA ⁺ (ICCV19) [17]	ResNet101	I+S	65.2	66.4
Zeng <i>et al.</i> (ICCV19) [43]	DenseNet	I+S'	63.3	64.3
CIAN (AAAI20) [10]	ResNet101	I+S	64.3	65.3
Zhang <i>et al.</i> (AAAI20) [44]	ResNet38	I	62.6	62.9
Luo <i>et al.</i> (AAAI20) [26]	ResNet101	I	64.5	64.6
Chang <i>et al.</i> (CVPR20) [3]	ResNet101	I	66.1	65.9
ICD (CVPR20) [9]	ResNet101	I+S	67.8	68.0
Araslanov <i>et al.</i> (CVPR20) [2]	ResNet38	I	62.7	64.3
SEAM (CVPR20) [36]	ResNet38	I	64.5	65.7
Zhang <i>et al.</i> (ECCV20) [47]	ResNet50	I+S	66.6	66.7
Sun <i>et al.</i> (ECCV20) [32]	ResNet101	I+S	66.2	66.9
CONTA (NeurIPS20) [45]	ResNet38	I	66.1	66.7
AuxSegNet* (Ours)	ResNet38	I+S	66.1	66.1
AuxSegNet (Ours)	ResNet38	I+S	69.0	68.6

Table 2. Performance comparison of WSSS methods in terms of mIoU(%) on the MS COCO *val* set.

Method	Backbone	Sup.	Val
SEC (CVPR16) [18]	VGG16	I+S	22.4
DSRG (CVPR18) [15]	VGG16	I+S	26.0
Wang <i>et al.</i> (IJCV20) [34]	VGG16	I	27.7
Luo <i>et al.</i> (AAAI20) [26]	VGG16	I	29.9
SEAM (CVPR20) [36]	ResNet38	I	31.9
CONTA (NeurIPS20) [45]	ResNet38	I	32.8
AuxSegNet (Ours)	ResNet38	I+S	33.9

background class) for semantic segmentation. This dataset has three subsets, *i.e.*, training (*train*), validation (*val*) and test with 1,464, 1,449 and 1,456 images, respectively. Following common practice, *e.g.*, [6, 18], we used additional data from [11] to construct an augmented dataset with 10,582 images for training. **MS COCO** contains 81 classes (including a background class). It has 80K training images and 40K validation images. Note that only image-level ground-truth labels from these benchmarks are used in the training process.

Evaluation metrics. As in previous works [17, 19, 38, 15, 1, 35], we used the mean Intersection-over-Union (mIoU) of all classes between the segmentation outputs and the ground-truth images to evaluate the performance of our method on the *val* and *test* sets of PASCAL VOC and the *val* set of MS COCO. We also used precision, recall and mIoU to evaluate the quality of the pseudo segmentation labels. The results on the PASCAL VOC *test* set were obtained from the official PASCAL VOC online evaluation server.

Implementation details. In our experiments, models were implemented in PyTorch. We use ResNet38 [39, 1] as the backbone network. For data augmentation, we used random

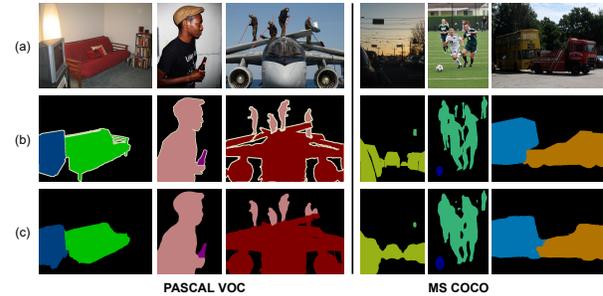


Figure 4. Qualitative segmentation results on the *val* sets of PASCAL VOC and MS COCO. (a) Inputs. (b) Ground-truth. (c) Our results.

scaling with a factor of ± 0.3 , random horizontal flipping, and random cropping to size 321×321 . The polynomial learning rate decay was chosen with an initial learning rate of 0.001 and a power of 0.9. We used the stochastic gradient descent (SGD) optimizer to train AuxSegNet with a batch size of 4. At inference, we use multi-scale testing and CRFs with hyper-parameters suggested in [6] for post-processing.

4.2. Comparison with State-of-the-arts

PASCAL VOC. We compared the segmentation performance of the proposed method with state-of-the-art WSSS approaches. Table 1 shows that the proposed method achieves mIoUs of 69.0% and 68.6% on the *val* and *test* sets, respectively. Our method outperforms the recent methods [10, 9, 32, 47] using off-the-shelf saliency maps by 4.7%, 1.2%, 2.8% and 2.4%, respectively, on the PASCAL VOC *val* set. The qualitative segmentation results on PASCAL VOC *val* set are shown in Figure 4 (left). Our segmentation results are shown to adapt well to different object scales and boundaries in various and challenging scenes.

MS COCO. To demonstrate the generalization ability of the proposed method, we also evaluated our method on the challenging MS COCO dataset. Table 2 shows segmentation results of recent methods on the *val* set. Our method achieves an mIoU of 33.9%, which is superior to state-of-the-art methods. Figure 4 (right) presents several examples of qualitative segmentation results, which indicate that our proposed method performs well in different complex scenes, such as small objects or multiple instances.

4.3. Ablative Analysis

Effect of auxiliary tasks. We compared results from the one-branch baseline model for semantic segmentation to several different variants: (i) baseline + cls: leveraging multi-label image classification, (ii) baseline + sal: leveraging saliency detection, and (iii) baseline + cls + sal: leveraging both image classification and saliency detection. Several conclusions can be drawn from Table 3. Firstly, the baseline performance with the single task of semantic segmentation is only 56.9%. Joint learning an auxiliary task

Table 3. Performance comparison of jointly learning different combinations of multiple tasks in terms of mIoU(%) on PASCAL VOC 2012 *val* set. *Seg.*, *Cls.*, and *Sal.* denote semantic segmentation, image classification, and saliency detection, respectively.

Config	Branches			mIoU
	Seg.	Cls.	Sal.	
Baseline	✓			56.9
AuxSegNet (w/ seg., cls.)	✓	✓		57.6
AuxSegNet (w/ seg., sal.)	✓		✓	59.8
AuxSegNet (w/ seg., cls., sal.)	✓	✓	✓	60.8

Table 4. Comparison of affinity learning with different settings in terms of mIoU(%) on PASCAL VOC 2012 *val* set. CT: cross-task.

Config	mIoU
AuxSegNet (multi-task baseline)	60.8
+ Seg. affinity with seg. constraint	61.5
+ CT affinity with seg. constraint	62.6
+ CT affinity with seg. and sal. constraints	64.1

Table 5. Segmentation performance of the proposed AuxSegNet after different training stages in terms of mIoU (%) on PASCAL VOC 2012 *val* set. Stage 0 denotes the training stage with the initial pseudo segmentation ground-truth without refinement.

	Stage0	Stage1	Stage2	Stage3	Stage4
mIoU	64.1	65.6	66.0	66.1	66.1

of either image classification or saliency detection both improve the segmentation performance significantly. In particular, learning saliency detection brings a larger performance gain of around 3%. Furthermore, leveraging both auxiliary tasks yields the best mIoU of 60.8% without using any post-processing. This indicates that these two related auxiliary tasks can improve the representational ability of the network to achieve more accurate segmentation predictions in the weakly supervised scenario.

Different settings for affinity learning. Table 4 shows ablation studies on the impact of different affinity learning settings on the segmentation performance. Without affinity learning, the segmentation mIoU is only 60.8%. The performance is improved to 61.5% by only learning segmentation affinity to refine segmentation predictions. Learning a cross-task affinity map which integrates both segmentation and saliency affinities brings a further performance boost of 1.1%. By forcing the cross-task affinity map to learn to refine both segmentation and saliency predictions, our model attains a significant improvement, reaching an mIoU of 64.1%. This demonstrates the positive effect of the multi-task constraints on learning pixel affinities to enhance weakly supervised segmentation performance.

Figure 5 presents several examples of the learned cross-task affinity maps for two selected points in each image. The affinity map for each pixel in an image is of the image size and it represents the similarity of this pixel to all pixels in the image. We can observe that, in the first row, the affin-

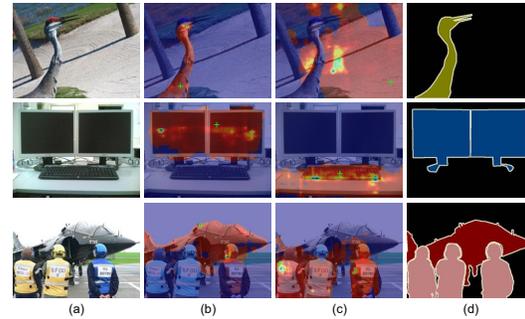


Figure 5. Visualization of the learned cross-task affinity maps of two selected points in the images on PASCAL VOC *train* set. (a) Inputs. (b)-(c) Two learned cross-task affinity maps for two points marked by the green crosses. (d) Segmentation ground-truth.

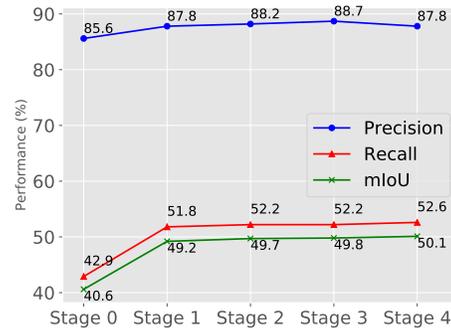


Figure 6. Evaluation of pseudo segmentation labels (PGT) for each training stage in terms of precision(%), recall(%) and mIoU(%) on PASCAL VOC 2012 *train* set.

ity map of the point labelled as “bird” highlights almost the entire object region although this point is far from the most discriminative object part “head”. Moreover, in the second row, the learned affinity map for the point belonging to the monitor activates most regions of the two monitor instances while it does not respond to the “keyboard” region which is similar in color in the background, and vice versa. In the third row, the affinity maps for the “airplane” and “person” points both present good boundaries.

Iterative improvements. To validate the effectiveness of the proposed iterative cross-task affinity learning, we evaluated the quality of the generated pseudo segmentation labels for each training stage. As shown in Figure 6, the precision, recall and mIoU of the initial pseudo segmentation labels generated by the CAM maps without refinement are 85.6%, 42.9% and 40.6%, respectively. After the first round of affinity learning, the updated pseudo labels for Stage 1 are significantly improved by 2.2%, 8.9% and 8.6% on these three metrics by using the refined CAM maps. Another round of affinity learning further boosts pseudo labels to 88.2%, 52.2% and 49.7% on the three metrics. In the following training stages, pseudo segmentation labels are slightly improved, and they are saturated at Stage 3 as we can observe that the precision starts to drop at Stage 4. As shown in Table 5, the segmentation performance presents

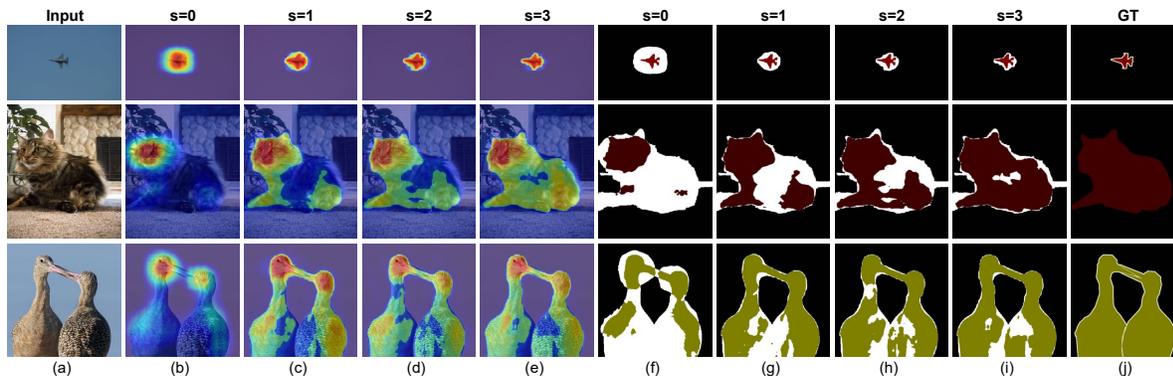


Figure 7. Visualization of CAM maps and pseudo segmentation labels with iterative improvements. (a) Inputs. (b) Initial CAM maps without refinement. (c)-(e) Refined CAM maps used to generate pseudo segmentation labels for Stage 1 to Stage 3. (f) Initial pseudo semantic labels for Stage 0. (g)-(i) Pseudo segmentation labels for Stage 1 to Stage 3. (j) Segmentation ground-truth. With the iterative cross-task affinity learning, the refined CAM maps become more complete with more accurate boundaries and the generated corresponding pseudo segmentation labels are closer to the ground-truth in terms of precision and recall.

Table 6. Segmentation results in terms of mIoU (%) using saliency models pre-trained in supervised or weakly supervised forms on PASCAL VOC 2012 *val* set. * denotes ‘without post-processing’.

Pre-trained saliency models	Baseline	Final*	Δ	Final
Zhang <i>et al.</i> (CVPR20) [46]	53.0	63.7	10.7	66.5
PoolNet (CVPR19) [24]	55.7	65.7	10.0	68.4
MINet (CVPR20) [27]	55.8	66.6	10.8	68.9
Ours with DSS [12]	56.9	66.1	9.2	69.0

consistent improvements as pseudo labels update after each training stage. Overall, the proposed AuxSegNet attains a reasonable improvement of 2% by using iterative label updates with the learned cross-task affinity, reaching the best mIoU of 66.1% without any post-processing.

To qualitatively evaluate the benefits of the proposed iterative affinity learning, Figure 7 presents several examples of CAM maps and the corresponding generated pseudo segmentation labels, and their iterative improvements along the training stages. As shown in Figure 7 (b), the CAM maps without affinity refinement for the initial training stage are either over-activated for small-scale objects, sparse for multiple instances or they only focus on the local discriminative object parts for large-scale objects. By refining CAM maps with the cross-task affinity learned at Stage 0, Figure 7 (c) shows that more object regions are activated for large-scale objects and the CAM maps for small-scale objects are more aligned to object boundaries. With more training stages, as shown in Figure 7 (d)-(e), the refined CAM maps become more integral with more accurate boundaries, which is attributed to the more reliable affinity learned with iteratively updated pseudo labels. The generated pseudo segmentation labels are shown to become progressively improved in Figure 7 (f)-(i), and they are close to the ground-truth labels.

Different pre-trained saliency models. To evaluate the sensitivity to the pre-trained saliency models, we conducted experiments using different pre-trained saliency models, *i.e.*, Zhang *et al.* [46], PoolNet [24] used in [32, 47], and

MINet [27], in which the first and the other two are in the weakly supervised and fully supervised forms, respectively. As shown in Table 6, our method achieves comparable results with these different saliency inputs, verifying the generalization ability of our method to different pre-trained saliency models. Moreover, with all these different pre-trained saliency models, our method can consistently produce significant performance improvements (see Δ) over the baseline, which further confirmed the effectiveness of the proposed method.

5. Conclusion

In this work, we proposed to leverage auxiliary tasks without requiring additional ground-truth annotations for WSSS. More specifically, we proposed AuxSegNet with a shared backbone and two auxiliary modules performing multi-label image classification and saliency detection to regularize the feature learning for the primary task of semantic segmentation. We also proposed to learn a cross-task pixel affinity map from saliency and segmentation feature maps. The learned cross-task affinity can be used to refine saliency predictions and CAM maps to provide improved pseudo labels for both tasks, which can further guide the network to learn more reliable pixel affinities and produce more accurate segmentation predictions. Iterative training procedures were thus conducted and realized progressive improvements on the segmentation performance. Extensive experiments on the challenging PASCAL VOC 2012 and MS COCO demonstrate the effectiveness of the proposed method and establish new state-of-the-art results.

Acknowledgment. This research is supported in part by Australian Research Council Grant DP150100294, DP150104251, DP200103223, Australian Medical Research Future Fund MR-FAI000085, SenseTime, the Early Career Scheme of the Research Grants Council (RGC) of the Hong Kong SAR under grant No. 26202321 and HKUST Startup Fund No. R9253.

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] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020.
- [3] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *CVPR*, 2020.
- [4] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *BMVC*, 2017.
- [5] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *CVPR*, 2016.
- [6] 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.
- [7] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.
- [9] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *CVPR*, 2020.
- [10] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *AAAI*, 2020.
- [11] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhansu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.
- [12] Q Hou, MM Cheng, X Hu, A Borji, Z Tu, and PHS Torr. Deeply supervised salient object detection with short connections. *PAMI*, 41(4):815–828, 2019.
- [13] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018.
- [14] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, 2018.
- [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.
- [16] Jaedong Hwang, Seohyun Kim, Jeany Son, and Bohyung Han. Weakly supervised instance segmentation by deep community learning. In *WACV*, 2021.
- [17] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *ICCV*, 2019.
- [18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.
- [19] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised segmentation using stochastic inference. In *CVPR*, 2019.
- [20] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018.
- [21] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *AAAI*, 2021.
- [22] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.
- [23] 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.
- [24] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.
- [25] Shikun Liu, Andrew Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. In *NeurIPS*, 2019.
- [26] Wenfeng Luo and Meng Yang. Learning saliency-free model with generic features for weakly-supervised semantic segmentation. In *AAAI*, 2020.
- [27] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, 2020.
- [28] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.
- [29] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *CVPR*, 2019.
- [30] Lv Sheng, Dan Xu, Wanli Ouyang, and Xiaogang Wang. Un-supervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam. In *ICCV*, 2019.
- [31] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *CVPR*, 2019.
- [32] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.
- [33] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *CVPR*, 2018.
- [34] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation by iterative affinity learning. *IJCV*, 128(6):1736–1749, 2020.
- [35] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018.
- [36] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020.

- [37] 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.
- [38] 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.
- [39] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.
- [40] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018.
- [41] Dan Xu, Andrea Vedaldi, and João F. Henriques. Moving slam: Fully unsupervised deep learning in non-rigid scenes. In *IROS*, 2021.
- [42] Lian Xu, Hao Xue, Mohammed Bennamoun, Farid Boussaid, and Ferdous Sohel. Atrous convolutional feature network for weakly supervised semantic segmentation. *Neurocomputing*, 421:115–126, 2021.
- [43] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019.
- [44] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *AAAI*, 2020.
- [45] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020.
- [46] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020.
- [47] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *ECCV*, 2020.
- [48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.