Neurocomputing 381 (2020) 20–28

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes

Xiang Wang[a,b,†], Huimin Ma[c,†,*], Shaodi You[d]

[a] Tencent Research, China
[b] Tsinghua University, China
[c] University of Science and Technology Beijing, China
[d] University of Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 8 April 2019
Revised 18 September 2019
Accepted 7 November 2019
Available online 13 November 2019

Communicated by Dr Ma Jiayi

Keywords:
Weak supervision
Semantic segmentation
Deep clustering
Autonomous driving

ABSTRACT

Weakly-supervised semantic segmentation (WSSS) using only tags can significantly ease the label costing, because full supervision needs pixel-level labeling. It is, however, a very challenging task because it is not straightforward to associate tags to visual appearance. Existing researches can only do tag-based WSSS on simple images, where only two or three tags exist in each image, and different images usually have different tags, such as the PASCAL VOC dataset. Therefore, it is easy to relate the tags to visual appearance and supervise the segmentation. However, real-world scenes are much more complex. Especially, the autonomous driving scenes usually contain nearly 20 tags in each image and those tags can repetitively appear from image to image, which means the existing simple image strategy does not work. In this paper, we propose to solve the problem by using region based deep clustering. The key idea is that, since each tagged object is repetitively appearing from image to image, it allows us to find the common appearance through region clustering, and particular deep neural network based clustering. Later, we relate the clustered region appearance to tags and utilize the tags to supervise the segmentation. Furthermore, regions found by clustering with weak supervision can be very noisy. We further propose a mechanic to improve and refine the supervision in an iterative manner. To our best knowledge, it is the first time that image tags weakly-supervised semantic segmentation can be applied in complex autonomous driving datasets with still images. Experimental results on the Cityscapes and CamVid datasets demonstrate the effectiveness of our method.

1. Introduction

Semantic segmentation aims to assign a semantic label to each pixel in images, it gives a full understanding of the scenes in images, thus can benefit a lot of applications, for example, autonomous driving [1–3]. However, semantic segmentation relies on a huge number of pixel-wise annotations, which is very costly and limits its application. Tags-based weakly-supervised semantic segmentation is a solution which can significantly reduce the labeling cost from pixel-level to a few tags per image.

In existing researches, tags-based methods have been explored on simple images, where only a few (usually two or three) tags exist in each image, and different images usually have different tags, such as the PASCAL VOC dataset. Therefore, it is relatively easy to relate tags to visual appearance and supervise the segmentation network.

However, real-world scenes are much more complex, especially the autonomous driving scenes, they usually contain nearly 20 tags in each image and those tags can repetitively appear from image to image, which means the existing simple image strategies [4–10] do not work. As shown in Fig. 2(a), in the PASCAL VOC dataset, only a few salient objects are presented, and a large part of regions are annotated as background. However, as in Fig. 2(b,c), in autonomous driving scenes, many objects are presented in one single image, some of them are even diverse and small. Facing the aforementioned difficulties, existing methods are not performing well in complex autonomous driving scenes. For example, the CCNN method [11] achieves mIoU at 35.6% on the PASCAL VOC dataset, but only obtains 7.2% on the Cityscapes dataset. To the best of our knowledge, image tags weakly-supervised semantic segmentation in complex still images, *e.g.*, the Cityscapes [1] and CamVid [12] datasets, has not be exploited.

In this paper, we aim to solve the tags-based semantic segmentation in autonomous driving scenes. The key idea is that, since

* Corresponding author at: University of Science and Technology Beijing, China.
E-mail addresses: andyxwang@tencent.com (X. Wang), mhmpub@ustb.edu.cn (H. Ma), s.you@uva.nl (S. You).
† Equal contribution.

https://doi.org/10.1016/j.neucom.2019.11.019
0925-2312/

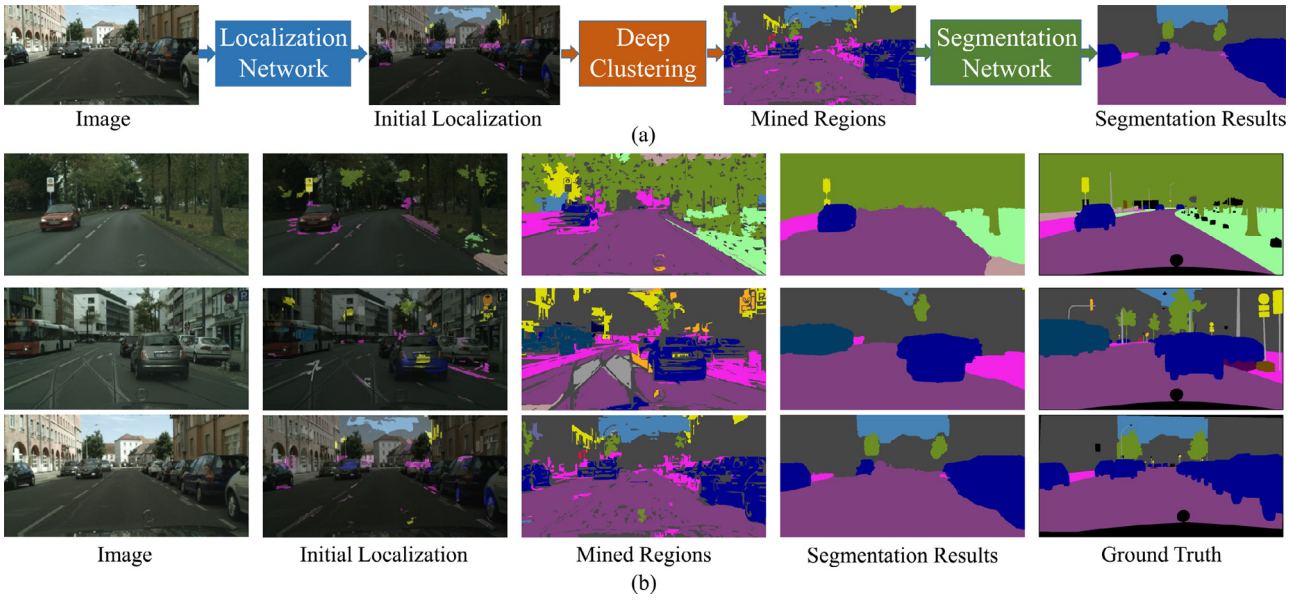


Fig. 1. (a) Pipeline of the proposed method. First, we apply the localization network trained from ImageNet dataset to get initial object localization. Second, with the initial object localization, we iteratively learn features of objects and cluster image regions to expand object regions. These regions are then used to supervise the segmentation network. (b) Some intermediate visual results. Starting from the very coarse initial localization, our method can produce quite satisfactory results.

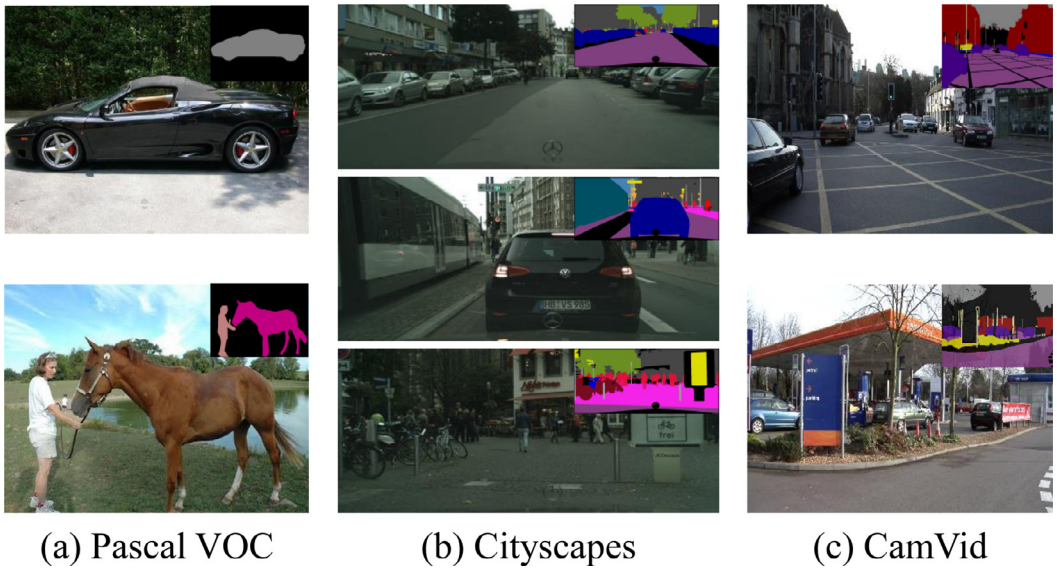


Fig. 2. Examples from the PASCAL VOC, Cityscapes and CamVid datasets. The top right is the corresponding thumbnail of annotation. In the PASCAL VOC dataset, the scene is simple with only a few objects presented in each image. While in the Cityscapes and the CamVid datasets, almost all object classes are presented in every single image simultaneously. Thus the class labels contain hardly any information for supervising networks.

each tagged object is repetitively appearing from image to image, it allows us to find the common appearance through region clustering, and particular deep neural network based clustering. Specifically, first, we take advantages of simple images, e.g., ImageNet dataset [13], and train a discriminative classifier to associate image tags with distinctive visual features. Then we apply the trained network to autonomous driving datasets and produce class activation map [14] as the initial localization. This localization provides us with discriminative regions of each object, as shown in the second column in Fig. 1(b).

Second, while complex autonomous driving scenes is more challenging, however, one import characteristic is that, in autonomous driving scenes, objects within the same class have more similarities, i.e., shared attributions, as all images in autonomous driving scenes have a very similar appearance. For example, in Fig. 3, cars appear in many images and their appearance does not vary much. In addition, objects are clustered, i.e., many objects appear in every single image, so they provide us with more training instances to learn the shared attributions of objects.

Motivated by this characteristic, we propose a novel iterative deep clustering method which learns shared attributions of objects and clusters image regions. The most straightforward idea is to directly cluster image regions. However, it is hard to design robust features to cluster them, besides, we cannot guarantee that each cluster is corresponding to each object class. So we propose to use the initial object localization as guidance and learn the shared attributions of objects from them. The learned model is then used to extract features of image regions and cluster them. This process is conducted iteratively, i.e., we further learn features from the clustered regions and then cluster them with more robust features. The

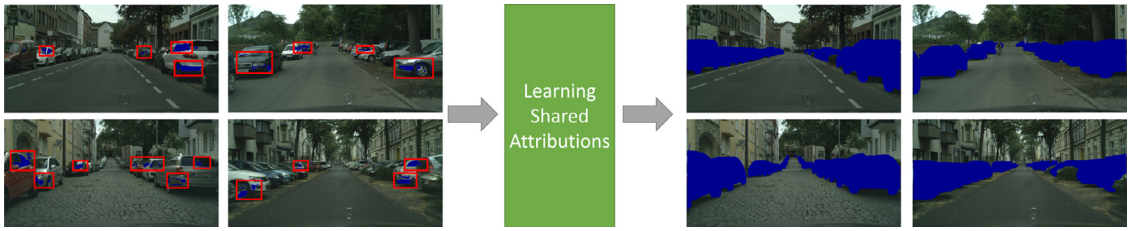


Fig. 3. In autonomous driving scenes, objects share more similarities, we can learn their shared contributions from coarse seeds regions to expand object region.

final clustered object regions are used for supervising a segmentation network.

Fig. 1 shows the pipeline and some visual examples of the initial localization and the produced result of our method. Though the initial localization is extremely coarse, our produced segmentation results achieve decent performance. The main contributions of our work are:

- We propose to learn discriminative visual features from simple images to produce initial object localization in complex autonomous driving scenes.
- We present an iterative deep clustering method which learns features and clusters image regions using initial object localization as guidance to expand object regions.
- To the best of our knowledge, this is the first solution which achieves single image weakly-supervised semantic segmentation in complex autonomous driving scenes with only image tags. We achieve better performance than the previous methods on the Cityscapes and the CamVid datasets.

2. Related work

2.1. Fully-supervised semantic segmentation

Semantic segmentation is a fundamental task in computer vision. In the past years, a large number of methods [15–22] have been proposed and achieved quite satisfactory results. Long et al. [15] propose fully convolutional networks (FCN) by introducing fully-convolutional layers to produce pixel-wise prediction of image semantic. Chen et al. [16,17] propose Deeplab network by introducing atrous convolution to enlarge the field of view of filters. Conditional Random Field (CRF) is applied as post-processing to improve localization performance. To consider more context information, Zhao et al. [18] present PSPNet which fuses different-region-based context with pyramid pooling module. Lin et al. [19] exploit all information from deep layers to shallow layers with a multi-path refinement network to produce segmentation results with high resolution. However, fully-supervised methods require a huge amount of pixel-level annotations, which is very time-consuming and thus limits its application.

2.2. Weakly-supervised semantic segmentation

Weakly-supervised semantic segmentation only requires much fewer annotations, for example, bounding box [23], scribbles [24], points [25] and image tags [4,8,10,11,26–32]. In this paper, we focus on supervision with image tags.

As only image tags, i.e., object classes, are available, previous weakly-supervised methods rely on classification network to locate objects. Many methods [4,11,30] solve weakly-supervised semantic segmentation as a Multi-Instance Learning (MIL) problem in which each image is taken as a package and contains at least one pixel of the known classes. By globally pooling the last feature map, the semantic segmentation problem is transformed to a classification task, and the segmentation results are produced from the last feature map. Kolesnikov and Lampert [8] put forward three kinds of losses based on initial object seeds produced by classification networks [14]. These losses aim to expand object regions from seeds and constrain them inside object boundary. Wei et al. [10] propose an adversarial erasing method which sequentially discovers new object regions from the erased images using classification networks.

These methods all rely on class labels to train classification networks and to produce discriminative regions of objects. However, when comes to complex autonomous driving scenes, as each image contains almost all classes, they cannot be used to train classification networks, so previous methods proposed for simple images will fail on complex autonomous driving scenes.

In [33], Saleh et al. propose weakly-supervised semantic segmentation method for urban scene by considering multiple background classes. However, it relies on the optical flow of videos as supervision to train a two-stream network, thus cannot be applied to scenes with still images. Among existing methods, [10,32,34] also adopt seed expansion strategy. Wei et al. use CAM to progressive mine discriminative regions from images, thus expanding object masks. Wang et al. iteratively mine robust common object features with a bottom-up and top-down framework. Huang et al. integrate seeded region growth (SGR) into semantic segmentation networks to expand seed regions. However, our method utilizes deep clustering framework to progressive learn robust clustering features and obtain better clusters, i.e., object masks. This is based on our observation that there exists much similarities in driving scenes, so we can cluster them to generate integral object masks.

2.3. Webly-supervised semantic segmentation

Webly-supervised semantic segmentation methods [35–38] use images queried from internet to train and locate object masks. These images are usually simple and easy to learn object masks, thus can be used as supervision of multi-class semantic segmentation. Our method also trains localization network from simple images and then transfers it to complex scenes. However, we have much differences with webly-supervised approaches. In webly-supervised methods, the simple images are usually taken as a supplementary supervision to boost typical weakly-supervised semantic segmentation networks. While in our work, the target domain, i.e., the driving scenes, cannot be used to train a localization network, and thus we propose to learn from simple images to associate image tags with distinctive visual features. In most webly-supervised methods, the transferred information is object masks, while ours is features.

3. The proposed framework

The motivation of our approach is that, in autonomous driving scenes, objects within the same class share more similarities, as all images have a very similar appearance. So we can learn the shared attributions of each class and use the learned features to cluster

Fig. 4. Produced object heatmap by classification network. (a) Directly applying trained network to whole image, the produced localization only focuses on few discriminative objects, e.g., road, while other objects are suppressed. (b) applying trained network to image patches. By processing images for each patch, the classification network can localize more diverse object regions, e.g., road, traffic sign, vegetation, sky, etc. Some non-exist objects, e.g., truck, bus, train, can be excluded as we have the image tags.

image regions. While directly clustering them is not feasible as it's hard to extract robust features and we can not guarantee that each cluster is corresponding to each object class. So we first produce initial object localization and use it as guidance to cluster image regions.

Our framework consists of three parts: (1) Initial object localization. We take advantages of simple images, e.g., ImageNet dataset, to train a discriminative classifier and then localize object seed regions with CAM method [14]. (2) Object seeds guided deep clustering. We learn shared attributions of each object class and use the learned attributions to cluster image regions to expand object regions. This process is conducted iteratively, i.e., learn shared attributions from the expanded object regions and get better clusters. (3) Weakly-supervised semantic segmentation. The produced object regions from part (2) are used as segmentation labels to train the segmentation network. Fig. 1 shows an illustration of the proposed framework and some intermediate visual results.

3.1. Initial object localization

One of the main challenges in weakly-supervised semantic segmentation is that only image tags are available, and no location information is provided. In simple scenes, such as the PASCAL VOC dataset, only one or a few objects are presented in images, so the tags associated with the image are highly distinguished and informative. However, in complex autonomous driving scenes, almost all object classes are presented in every single image and thus the class labels contain hardly any information to localize objects. To address this issue, we propose to take advantages of simple images, such as ImageNet dataset [13], to train a discriminative classifier which associates the image tags with distinctive visual features.

To be specific, first, we select classes from the ImageNet dataset which are corresponding to our dataset, and train a classification network with the VGG16 network [39]. With the trained network, we can then produce object heatmap using the Class Activation Map method (CAM) [14].

However, in autonomous driving scenes, objects are clustered, i.e., a large number of objects will appear in every single image. If directly apply the trained network, it is difficult to produce discriminative regions of each object. Some salient objects will affect other non-salient objects, so the produced localization only focuses on a part of discriminative objects. In addition, the size of objects is varied due to changes in distance, objects with small size are hard to locate when taking the whole image as input. Fig. 4(a) shows an example. When directly applied to the whole image, only the regions of road (the top left corner image) are emphasized, while other objects are suppressed. To address this problem, we propose to segment images into patches, so that the classification network can better handle discriminative objects and small objects. For the Cityscapes dataset, the original resolution is 2048×1024, we segment it into 8 patches (4×2), and for the CamVid dataset [12], the original resolution is 960×720, we segment it into 12 patches (4×3), so that each patch is square to prevent deformation. Fig. 4(b) shows an example when applying the classification network to image patches. We can see that the produced heatmap can localize more diverse objects. With these heatmaps, we then segment image into superpixel regions [40] and select regions with maximum object heatmap as initial object seeds.

3.2. Object seeds guided deep clustering

In autonomous driving scenes, one important characteristic is that, in autonomous driving scenes, objects within the same class have more similarities, i.e., shared attributions, as all images in autonomous driving scenes have a very similar appearance. In addition, objects are clustered, i.e., many objects appear in every single image, so they provide us with more training instances to learn the shared attributions of objects. Motivated by this, we propose to cluster image regions to generate region masks of each object. The most straightforward idea is to direct cluster these objects. However, it is hard to design robust features to cluster them, besides, we can not guarantee that each cluster is corresponding to each object class. To address these issues, in this paper, we argue that the initial object seeds provide us with significant information about objects, so we propose a deep clustering method with the initial object seeds as guidance.

In detail, we train a neural network using the initial object seeds and then extract features of other regions with the trained network. First, we segment images into superpixel regions [40], superpixels within the initial object seeds are marked with corresponding class labels and other superpixels are labeled as unknown. We then use these labeled superpixels as supervision to train a region classification network. With the trained network, we extract features of each region and use these features to repredict new labels for all superpixel regions. We name our method as *deep clustering* as we use the learned features to cluster regions and our objective is to minimize the variance within the same class and maximize the variance among different classes. These processes are conducted iteratively, i.e., we further use the new predicted labels as supervision to optimize the network to progressively mine robust features and object regions.

Formally, given a set of M training images $\mathcal{I} = \{I_i\}_{i=1}^{M}$, we segment image I_i into N_i superpixel regions to obtain $\mathcal{I}^{\mathcal{R}} = \{I_{i,j}^{R}\}_{i=1,j=1}^{M,N_i}$. With the produced initial object seeds $\mathcal{S} = \{S_{i,j}\}_{i=1,j=1}^{M,N_i}$, our goal is to optimize the network θ to mine robust features to represent shared attributes of objects, and thus to predict accurate object la-

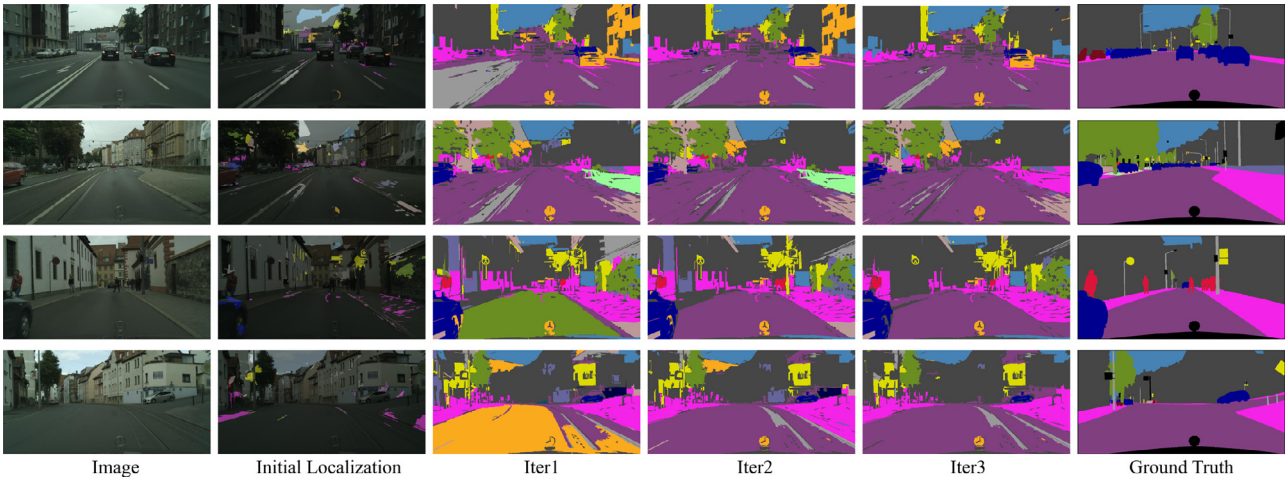


Fig. 5. Visual examples of iterative deep clustering. Starting from very coarse and inaccurate initial localization, the object seeds guided deep clustering can progressively cluster and expand object regions.

bels y by solving

$$\arg\min_{y,\theta} \mathcal{L}(y, \theta | I^R). \tag{1}$$

We solve Eq. (1) by fixing one of the parameters $\{y, \theta\}$ and decompose it into two alternating steps:

$$\arg\min_{\theta} \mathcal{L}(\theta | y, I^R), \tag{2}$$

$$\arg\min_{y} \mathcal{L}(y | \theta, I^R). \tag{3}$$

In Eq. (2), we fix the region labels to optimize the network parameter θ. And in Eq. (3), with the trained network parameter θ, we predict class labels of given image regions I^R. These two steps are iteratively optimized to progressively mine shared attributes of objects in each class and cluster object regions.

We realize the region classification network with the Mask-based Fast R-CNN framework [41], which can extracts features of irregular superpixel regions efficiently. Our goal is to learn high-level features and use them to cluster regions. To achieve better performance, we require the learned features to have large variance among different classes and small variance within the same class. To this end, we introduce two kinds of losses to train the network. The first loss is cross-entropy loss, which encourages large variance among different classes, and it is defined as:

$$\mathcal{L}_1 = -\sum_{i,j,c} S_{i,j}(c) log(f_c(I^R_{i,j}|\theta)), \tag{4}$$

where $S_{i,j}(c)$ is the label of region $I^R_{i,j}$. $S_{i,j}(c) = 1$ if region $I^R_{i,j}$ belongs to class c, otherwise $S_{i,j}(c) = 0$. $f_c(I^R_{i,j}|\theta)$ denotes the classification score of region $I^R_{i,j}$ being predicted as class c.

The second loss is the center loss [42], which aims to produce small variance within the same class:

$$\mathcal{L}_2 = \frac{1}{2}\sum_{i,j} ||\boldsymbol{x}_{i,j} - \boldsymbol{m}_{y_{i,j}}||_2^2, \tag{5}$$

where $\boldsymbol{x}_{i,j}$ is the features of region $I^R_{i,j}$, $\boldsymbol{m}_{y_{i,j}}$ is the center of features of class $y_{i,j}$.

Finally, we optimize Eq. (2) with the following joint supervision:

$$\mathcal{L}_C = \mathcal{L}_1 + \lambda \mathcal{L}_2, \tag{6}$$

we set $\lambda = 0.001$ based on our experiments and previous work [42].

Fig. 5 shows some examples of our deep clustering. With the extreme coarse initial localization, our deep clustering can produce finer annotation, and with the iterative optimization, the performance improves gradually.

3.3. Weakly-supervised semantic segmentation

With the object seeds guided deep clustering, object regions are clustered to corresponding classes and thus the object regions are expanded. With these object regions, we then take them as segmentation labels to train the segmentation network. The training is the same as any fully-supervised semantic segmentation network. In this paper, we utilize the popular DeepLab-LargeFOV model [43] as the basic network and use the cross-entropy loss to optimize the network. The final trained network is then used for inference.

4. Experiments

4.1. Setup

We evaluate the proposed method on two challenging datasets: Cityscapes [1] and CamVid [12]. The Cityscapes dataset contains urban scenes of 50 cities, and provides 5000 images with fine annotations (*train*: 2975, *val*: 500, *test*: 1525). There are 30 classes and 7 categories in this dataset, and for public evaluation, 19 classes are considered while leaving others as void. The CamVid dataset contains 701 annotated images for semantic segmentation (*train*: 367, *val*: 101, *test*: 233). There are 32 semantic classes in this datasets, following previous works [33,44], only 11 classes are evaluated. In this paper, we train our method in the training set and evaluate it on the val and test sets in terms of intersection-over-union averaged on all classes (mIoU Class). For Cityscapes dataset, intersection-over-union averaged on categories (mIoU Category) is also evaluated.

4.2. Implementation details

Initial object seeds. We select images from ImageNet dataset with same classes with Cityscapes or CamVid datasets to train the classification network. The network we used is VGG16 [39].

Object seeds guided deep clustering. We realize the object seeds guided deep clustering with a region classification network supervised by cross-entropy loss and center loss. Directly training the network with joint supervision will make the prediction bias to

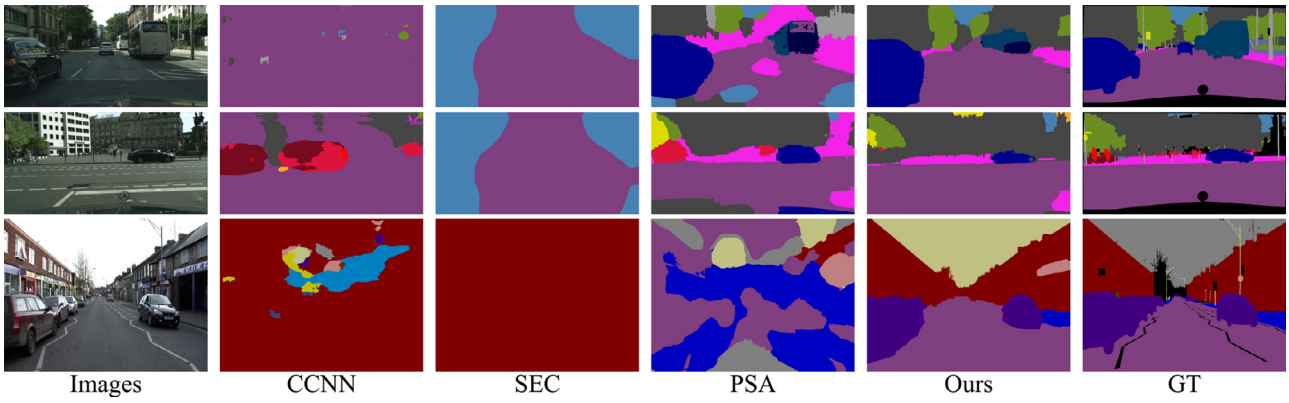


Fig. 6. Qualitative segmentation results on Cityscapes and CamVid datasets.

Table 1
Results on the Cityscapes *val* set.

Methods	Supervision	mIoU Class	mIoU Cat.
CCNN [ICCV'15]	Image tags	7.3	16.3
SEC [ECCV'16]	Image tags	2.3	7.1
PSA [CVPR'18]	Image tags	21.6	39.0
Ours	Image tags	**24.2**	**50.2**

Table 2
Results on the Cityscapes *test* set.

Methods	Supervision	mIoU Class	mIoU Cat.
CCNN [ICCV'15]	Image tags	7.2	17.2
SEC [ECCV'16]	Image tags	2.4	7.2
BBF [ICCV'17]	Tags & videos	24.9	47.2
PSA [CVPR'18]	Image tags	21.2	40.2
Ours	Image tags	**24.9**	**53.7**

Table 3
Results on CamVid *val* set.

Methods	Supervision	mIoU
CCNN [ICCV'15]	Image tags	2.9
PSA [CVPR'18]	Image tags	11.0
Ours	Image tags	**23.5**

Table 4
Results on CamVid *test* set.

Methods	Supervision	mIoU
CCNN [ICCV'15]	Image tags	2.4
SEC [ECCV'16]	Image tags	2.5
BBF [ICCV'17]	Image tags & videos	29.7
PSA [CVPR'18]	Image tags	15.5
Ours	Image tags	**30.4**

objects with larger regions, for example, road and building, as the center loss will get a very low value in that case. So we first train the network with only cross-entropy loss for 40,000 iterations, then we continue to finetune it with both losses for 5000 iterations.

Weakly-supervised semantic segmentation network. We use the DeepLab-LargeFOV [43] pre-trained on ImageNet as basic network of our segmentation network.

All the networks are realized and trained on Caffe framework [45]. All code will be released and more details can be found in it.

4.3. Evaluation and comparison

We evaluate our method on Cityscapes [1] and CamVid [12] datasets, and compare with previous methods. To our best knowledge, weakly-supervised semantic segmentation in complex scenes with only image tags, such as Cityscapes and CamVid datasets with still images, has not be exploited before. For comparison, we implement previous weakly-supervised methods on Cityscapes and CamVid datasets. As only a few methods have code released, here we compare with CCCN [11], SEC [8] and PSA [46]. These methods are proposed for simple scenes, i.e., the PASCAL VOC dataset. We also compare with a weakly-supervised method based on videos: BBF [33]. For CCCN [11], it's an end-to-end method, so we directly apply it to autonomous driving scenes. For SEC [8] and PSA [46], they rely on initial localization, but their methods cannot get initial localization on autonomous driving scenes, here, we use localization generated by our method. The results are shown inTables 1–4. Our method outperforms all other methods. For CCNN method, it is based on multi-instance learning, so it mainly focuses on objects with large size, such as road and building. For SEC method, it has three kinds of constrain to train the network, these constrain is not applicable in complex scenes, it's output has a very strong bias, such as road and sky in Cityscapes, and building in CamVid. For PSA, it is based on the initial localization of our method, and then learn affinity to refine it, so it achieves a relative decent performance. For BBF, it also uses images from ImageNet dataset, in addition, it also uses optical flow in videos, however, our method still performs better than it with only still images. Some qualitative results are shown in Fig. 6.

4.4. Ablation studies

We conduct experiments to evaluate the effectiveness of our method, all results are evaluated on the Cityscapes dataset.

4.4.1. Localization from image patches

We evaluate the effectiveness of generating initial localization from image patches by comparing with directly applied to whole images. Fig. 7 shows some examples when we apply the trained network to whole images and image patches. We can see that when directly applied to whole images, the localization heatmap is coarse and diffused, and when applied to image patches, the heatmap can better locate discriminative regions of objects. We also list the localization accuracy in terms of IoU in Table 5. These results demonstrate the effectiveness of generating initial localization from image patches. Our method can prevent the impact that salient objects suppress the heatmap of other objects, and we can better handle objects with small size, for example, light, sign and person. In addition, if we directly generate from whole images, we

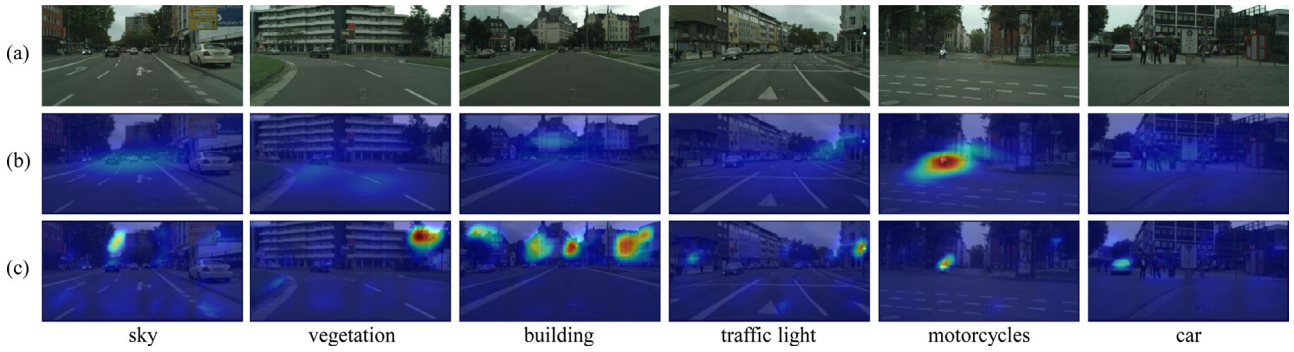


Fig. 7. Comparison of generating initial localization from (b) whole images and (c) image patches. When directly applied to whole images, the produce heatmap only focus on a few salient objects, and objects with small size are hard to locate. While applied to each patch of images, these issues can be better handled.

Table 5
Evaluate the effectiveness of generating initial object seeds from image patches.

	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Whole image	3.2	5.0	0.6	0.3	0.2	0.0	3.4	0.2	0.3	0.5	0.0	1.0	1.6	0.5	3.3	7.4	15.0	1.4	0.0	2.3
Image patches	0.7	3.7	5.0	1.3	1.2	1.2	10.3	9.3	4.4	1.8	4.9	3.9	1.0	4.4	4.8	7.6	4.6	3.9	0.4	**3.9**

Table 6
Evaluate the effectiveness of deep clustering.

Methods	mIoU Class	mIoU Cat.
Seeds	3.9	4.7
Deep Clustering-1	16.5	21.1
Deep Clustering-2	19.1	37.7
Deep Clustering-3	19.6	38.3

Table 7
Evaluate the effectiveness of joint supervision with center loss.

	Without center loss		With center loss	
	mIoU Class	mIoU Cat.	mIoU Class	mIoU Cat.
DC-1	16.2	19.3	16.5	21.1
DC-2	18.6	36.7	19.1	37.7
DC-3	19.6	37.9	19.6	38.3

cannot obtain localization for classes of pole, sky and bike, while our approach can provide significant localization information for all classes.

4.4.2. Iterative deep clustering

To evaluate the effectiveness of the iterative deep clustering, we list the intermediate results in Table 6. The initial object seeds provide very coarse localization, only 3.9% in mIoU Class and 4.7% in mIoU Category. However, these seeds give us significant knowledge about objects, with the object seeds guided deep clustering, more object regions are clustered to correct classes, and the performance is increased by a large margin and achieves 16.5% in mIoU Class and 21.1% in mIoU Category. In the later iterations, the object regions are gradually corrected and the performance improves progressively.

4.4.3. Joint supervision with center loss

To evaluate the effectiveness of the joint supervision with the center loss, we compare the performance with the method using only cross-entropy loss. Table 7 shows the results on the Cityscapes dataset. With the help of center loss, we can achieve a relatively higher performance.

4.4.4. Results on the PASCAL VOC 2012 dataset

Our method is also applicable to simple scenes, such as the PASCAL VOC 2012 dataset. For initial localization, following previous methods, we directly train classification network with images from the PASCAL VOC 2012 dataset. Table 8 shows the comparison with up-to-date weakly-supervised methods: CCNN [11], SEC [8], STC [35], SPN [28], AE-PSL [10], LCEM [31], MCOF [34] and DSRG [32]. Though we aim to solve weakly-supervised semantic segmentation in complex autonomous driving scenes, on the simple PASCAL VOC 2012 dataset, our approach is better than most previous methods. However, on the PASCAL VOC 2012 dataset, the similarities between images are weaker than in the driving scenes, so our method is inferior to recent state-of-the-art approaches which are designed for the PASCAL VOC dataset.

Table 8
Comparison on the PASCAL VOC 2012 dataset.

Methods	*val* set	*test* set
CCNN [11] [ICCV'15]	35.3	35.6
SEC [8] [ECCV'16]	50.7	51.7
STC [35] [PAMI'17]	49.8	51.2
SPN [28] [AAAI'17]	50.2	46.9
AE-PSL [10] [CVPR'17]	55.0	55.7
LCEM [31] [NEUCOM'18]	45.4	46.0
MCOF [34] [CVPR'18]	56.2	57.6
DSRG [32] [CVPR'18]	59.0	60.4
Ours	**55.6**	**57.2**

Table 9
Results with other segmentation networks on the Cityscapes *val* set.

Segmentation Networks	mIoU class	mIoU cat.
FCN [15]	22.3	49.4
DRN-D-105 [47]	23.7	**50.6**
DeepLab-LargeFOV [43]	**24.2**	50.2

4.4.5. Results with other segmentation networks

We further conduct experiments using different segmentation networks: FCN [15], DRN-D-105 [47] and DeepLab-LargeFOV [16]. We use the synthesized segmentation labels from the proposed deep clustering as pseudo-supervision to train these networks, and evaluate them on the Cityscapes validation set. Table 9 shows the results. As the proposed approach is a unified weakly-supervised learning framework, we can take advantages of all existing segmentation networks in different situations.

Table 10
IoU of all classes on Cityscapes *val* set.

	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Cityscapes	57.1	19.3	61.5	0.0	1.3	2.8	3.4	10.6	58.5	6.2	50.4	35.9	0.0	63.4	4.4	21.9	5.0	19.5	38.2	**24.2**

4.5. Failure cases

The proposed method solves weakly-supervised semantic segmentation in complex driving scenes and performs better than previous method which uses additional video information. While driving scenes are very challenging, there are some failure cases worth further researching. Table 10 shows the IoU for all classes in the Cityscapes validation set. The proposed method achieves relative good performance on some large or common objects, such as road, building, vegetation, sky, person, car and bike. The reason is that these classes have large area or more instances, so our clustering method can effectively cluster them and obtain better object masks. However, for some classes that are rare to appear, such as wall and fence, the performance is still quite low, as they have less samples to train the deep clustering. In addition, for the rider class, as it is very confusing with person and bike, it is also difficult to segment. In the future work, we will explore solution for these hard classes.

5. Conclusion

In this paper, we solve weakly-supervised semantic segmentation in complex autonomous driving scenes. To localize objects, we propose to learn discriminative visual features from simple images, and produce initial localization with the learned features in autonomous driving scenes. Using the initial localization as object seeds, an object seeds guided deep clustering method is proposed to iteratively learn shared attributions of objects of each class and to expand object regions. These object regions are then used as supervision to train the segmentation network. Experimental results on the Cityscapes and CamVid datasets demonstrate that our method achieves decent performance, and we also outperform the previous state-of-the-art weakly-supervised method which used addition optical flow in videos as supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by National Key Basic Research Program of China (No. 2016YFB0100900), Beijing Science and Technology Planning Project (No. Z191100007419001) and National Natural Science Foundation of China (No. 61773231).

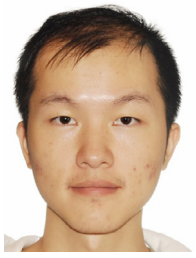
References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[2] Y. Zhang, H. Chen, Y. He, M. Ye, X. Cai, D. Zhang, Road segmentation for all-day outdoor robot navigation, Neurocomputing 314 (2018) 316–325.
[3] G. Tian, L. Liu, J. Ri, Y. Liu, Y. Sun, Objectfusion: an object detection and segmentation framework with rgb-d slam and convolutional neural networks, Neurocomputing 345 (2019) 3–14.
[4] D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning, in: Proceedings of International Conference on Learning Representations (ICLR) Workshop, 2015.
[5] P.O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[6] F. Saleh, M.S.A. Akbarian, M. Salzmann, L. Petersson, S. Gould, J.M. Alvarez, Built-in foreground/background prior for weakly-supervised semantic segmentation, in: Proceedings of European Conference on Computer Vision, ECCV, 2016.
[7] G. Papandreou, L.-C. Chen, K. Murphy, A.L. Yuille, Weakly-and Semi-supervised Learning of a dcnn for Semantic Image Segmentation, Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015.
[8] A. Kolesnikov, C.H. Lampert, Seed, expand and constrain: three principles for weakly-supervised image segmentation, in: Proceedings of European Conference on Computer Vision, ECCV, 2016.
[9] X. Qi, Z. Liu, J. Shi, H. Zhao, J. Jia, Augmented feedback in semantic segmentation under image level supervision, in: Proceedings of European Conference on Computer Vision, ECCV, Springer, 2016.
[10] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: s simple classification to semantic segmentation approach, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[11] D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: Proceedings of International Conference on Computer Vision, ICCV, 2015.
[12] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high-definition ground truth database, Pattern Recognition Letters 30.2 (2009) 88–97.
[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115.3 (2015) 211–252.
[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[15] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic Image Segmentation with Deep Convolutional Nets And fully Connected CRFS, International Conference on Learning Representations, ICLR (2015).
[17] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, in: Proceedings of European Conference on Computer Vision, ECCV, 2018.
[18] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[19] G. Lin, A. Milan, C. Shen, I.D. Reid, Refinenet: multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[20] Z. Wu, C. Shen, A. d Hengel, Wider or Deeper: Revisiting the Resnet Model for Visual Recognition, Pattern Recognition 90 (2019) 119–133.
[21] F. Shen, G. Zeng, Semantic image segmentation via guidance of image classification, Neurocomputing 330 (2019) 259–266.
[22] Z. Jiang, Y. Yuan, Q. Wang, Contour-aware network for semantic segmentation via adaptive depth, Neurocomputing 284 (2018) 27–35.
[23] J. Dai, K. He, J. Sun, Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: Proceedings of International Conference on Computer Vision, ICCV, 2015.
[24] D. Lin, J. Dai, J. Jia, K. He, J. Sun, Scribblesup: scribble-supervised convolutional networks for semantic segmentation, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[25] A. Bearman, O. Russakovsky, V. Ferrari, L. Fei-Fei, Whats the point: semantic segmentation with point supervision, in: Proceedings of European Conference on Computer Vision, ECCV, 2016.
[26] A. Chaudhry, P.K. Dokania, P.H. Torr, Discovering class-specific pixels for weakly-supervised semantic segmentation, in: Proceedings of British Machine Vision Conference, BMVC, 2017.
[27] A. Diba, V. Sharma, A.M. Pazandeh, H. Pirsiavash, L. Van Gool, Weakly supervised cascaded convolutional networks., in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[28] S. Kwak, S. Hong, B. Han, et al., Weakly supervised semantic segmentation using superpixel pooling network., in: Proceedings of AAAI, 2017.
[29] T. Durand, T. Mordan, N. Thome, M. Cord, Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation, in: Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[30] A. Roy, S. Todorovic, Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[31] Y. Li, Y. Liu, G. Liu, D. Zhai, M. Guo, Weakly supervised semantic segmentation based on em algorithm with localization clues, Neurocomputing 275 (2018) 2574–2587.

[32] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7014–7023.
[33] F.S. Saleh, M.S. Aliakbarian, M. Salzmann, L. Petersson, J.M. Alvarez, Bringing background into the foreground: making all classes equal in weakly-supervised video semantic segmentation, in: Proceedings of International Conference on Computer Vision, ICCV, 2017.
[34] X. Wang, S. You, X. Li, H. Ma, Weakly-supervised semantic segmentation by iteratively mining common object features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1354–1362.
[35] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, STC: a simple to complex framework for weakly-supervised semantic segmentation, Proceedings of IEEE TPAMI (2016).
[36] S. Hong, D. Yeo, S. Kwak, H. Lee, B. Han, Weakly supervised semantic segmentation using web-crawled videos, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[37] B. Jin, M.V. Ortiz Segovia, S. Susstrunk, Webly supervised semantic segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[38] T. Shen, G. Lin, C. Shen, I. Reid, Bootstrapping the performance of webly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1363–1371.
[39] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, International Conference on Learning Representations, ICLR (2015).
[40] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient Graph-based Image Segmentation, Int. J. Comput. Vis. 59.2 (2004) 167–181.
[41] X. Wang, H. Ma, X. Chen, S. You, Edge preserving and multi-scale contextual neural network for salient object detection, IEEE Trans. Image Process. 27.1 (2018) 121–134.
[42] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proceedings of European Conference on Computer Vision, ECCV, 2016.
[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[44] A. Kundu, V. Vineet, V. Koltun, Feature space optimization for semantic video segmentation, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[45] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of ACM MM, 2014.
[46] J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4981–4990.
[47] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 472–480.

Xiang Wang received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2014 and 2019, respectively. He is currently a researcher at Tencent. His research interests are computer vision and machine learning, with particular interests in salient object detection, semantic segmentation and weakly-supervised learning. He serves as a reviewer for TIP, CVPR, ICCV and AAAI.

Huimin Ma received the M.S. and Ph.D. degrees in mechanical electronic engineering from the Beijing Institute of Technology, Beijing, China, in 1998 and 2001, respectively. She is currently the Vice Dean of the Institute of Artificial Intelligence, University of Science and Technology Beijing. She was a Visiting Scholar with University of Pittsburgh in 2011. She is also the Secretary General of China Society of Image and Graphics. Her research and teaching interests include 3D object recognition and tracking, system modeling and simulation, and psychological base of image cognition.



Shaodi You received the bachelor's degree from Tsinghua University, China, in 2009, the. M.E. and Ph.D. degrees from The University of Tokyo, Japan, in 2015 and 2012. He is currently an Assistant Professor at University of Amsterdam (UvA), Netherlands. His research interests are physics based vision, nonrigid 3D geometry and perception and learning based vision. He is currently the Chair of IEEE Computer Society, Australian Capital Territory Section, Australia. He is the Program Chair of ICCV2017 Joint Workshop on Physics Based Vision meets Deep Learning. He serves as a Reviewer for TPAMI, IJCV, TIP, CVPR, ICCV, and SIGGRAPH.