Saliency guided deep network for weakly-supervised image segmentation

Fengdong Sun[a], Wenhui Li[a,**]

[a]College of Computer Science and Technology, Jilin University, Changchun, 130012, China

arXiv:1810.08378v1 [cs.CV] 19 Oct 2018

ABSTRACT

Weakly-supervised image segmentation is an important task in computer vision. A key problem is how to obtain high quality objects location from image-level category. Classification activation mapping is a common method which can be used to generate high-precise object location cues. However these location cues are generally very sparse and small such that they can not provide effective information for image segmentation. In this paper, we propose a saliency guided image segmentation network to resolve this problem. We employ a self-attention saliency method to generate subtle saliency maps, and render the location cues grow as seeds by seeded region growing method to expand pixel-level labels extent. In the process of seeds growing, we use the saliency values to weight the similarity between pixels to control the growing. Therefore saliency information could help generate discriminative object regions, and the effects of wrong salient pixels can be suppressed efficiently. Experimental results on a common segmentation dataset PASCAL VOC2012 demonstrate the effectiveness of our method.

1. Introduction

Recently, computer vision research has a prominent progress and achieves excellent performance. Many tasks in computer vision field need plenty pixel-level annotations to guarantee the accuracy of the corresponding solutions, such as scene understanding (Wang et al., 2017b) and instance segmentations (Wu et al., 2018a). The pixel-level annotations indicate that each pixel in the ground truth has a label referring to its category. However, it is very difficult to obtain such pixel-level annotation datasets, because this kind of annotations is time consuming and requires substantial nancial investments. The process of labeling a pixel-level ground truth generally consumes a subject several minutes. On the contrast, weakly-labeled visual data, which only indicate the categories included by images but do not provide the locations of these categories, can be obtained in a relatively fast and cheap manner. Therefore, it is important and meaningful to generate pixel-level annotation data using weakly-labeled images, i.e. weakly-supervised semantic segmentation(Wang et al., 2015).

In this paper, we focus on conducting pixel-labeled segmentation using weakly-labeled data. However, there is a large per-

formance gap between weakly and fully supervised image semantic segmentation (Wu and Wang, 2018; Wu et al., 2018b). A key problem is how to infer the objects locations according to image-level categories. (Qi et al., 2016) used objectness proposal information to guide a object localization network to generate location cues, then aggregating these cues to help semantic segmentation. Although there are lots of helpful information contained in these aggregated location cues. Meanwhile many interference informations are mixed into them. These interferences are difficult to distinguish and eliminate under weakly-supervision such that they may effect the accuracy of object localization and image semantic segmentation. (Kolesnikov and Lampert, 2016) employed a classification network to retrieve objects location cues based on classification activation maps. These location cues, which consist of some discriminative regions, are very reliable and robust that could be used to improve the performance of segmentation tasks. Therefore, (Kolesnikov and Lampert, 2016) used these location cues to train a semantic segmentation network immediately. However, the discriminative regions in location cues are too small and sparse that they do not have enough ability to tune the entire network(Wu et al., 2018d).

For obtaining complete objects location from small and sparse cues, saliency detection methods are developed to enhance the performance of weakly-supervised segmentation.

[**]Corresponding author
   e-mail: liwh@jlu.edu.cn (Wenhui Li)

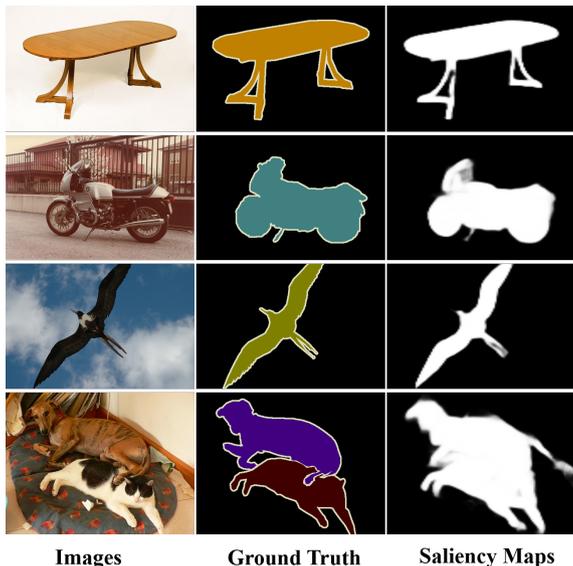**Images**   **Ground Truth**   **Saliency Maps**

Fig. 1. Some images in PACAL VOC 2012 validation set with their ground truths and saliency maps.

Saliency information of an image uses a saliency map to indicate the regions that most attract human beings' attentions. The saliency map can used to segment the input image into foreground and background. The salient foregrounds have a clear boundary and generally contain several salient objects, therefore could be utilized to generate objects locations from precise and reliable cues as Figure 1 shown. (Joon Oh et al., 2017) propose to utilize saliency to assist semantic segmentation. However, they assign salient regions a category randomly picked from image-level labels initially, and these salient regions will be assigned to a category if the category's seed touching the salient regions afterwards. This method heavily depends on the precision of the saliency methods and may produce sub-optimal results if a salient regions just have an incorrect pixel touching a category seed.

To address the aforementioned problem, we propose a novel method called saliency guided weakly-supervised segmentation network which utilize saliency information as a guidance to generate robust objects locations from sparse cues to help image segmentation. Firstly, we use a weakly objection localization network to generate locations seeds from image-level category. These seeds have high confidence and precision that could be regard as ground truths. Secondly, to resolve the sparse and small issues of location seeds, we propose a novel method called saliency guided seeded region growing. The saliency information we used is from a self-attention saliency network which utilize image inherent cues, i.e. self-attention, to generate stage-wise refined saliency maps (Sun et al., 2018). We use the saliency detection method as a black box in this paper and immediately use the final saliency maps to guide the process of seeded region growing from location seeds. To alleviate the effect incorrect saliency results caused, we do not assign the salient region with a same label. We use the saliency values of pixels to generate saliency weights to control the process of seeded region growing. Therefore, a pixel not in a salient regions have possibility to get corresponding label, and a wrong salient pixel may be not grew by the location seeds. The saliency guidance can make the pixels with same saliency property easy to have the same label. At last, we integrate these cues into a network for weakly-supervised image segmentation. Experimental results demonstrate that our method outperforms several methods on a common PASCAL VOC2012 dataset.

In summary, the contributions of this paper are as followings:

1. We integrate weakly objects localization, saliency detection and saliency guided seeded region growing into a deep network framework for weakly-supervised segmentation.

2. We propose a seeded region growing method with saliency guidance to expand the location generated by classification activation maps, therefore enriching pixel-level segmentation information.

3. Experiments on a common dataset PASCAL VOC2012 demonstrate our method has a better performance than 11 existing algorithms.

2. Related Work

2.1. Saliency detection methods

Many saliency detection methods are exploited for exact foreground segmentation recently(Wang et al., 2018). In general, these methods can be divide into two categories: traditional methods and deep learning methods. Many researches demonstrate that deep learning methods have a significant improvement in accuracy of saliency detection, and various different modules are exploited to enhance the performance of deep saliency networks. (Wang et al., 2017a) propose a stagewise refinement model to refine the saliency maps. A coarse prediction map is generated by the model firstly. Then the a refinement structure is used to refine the coarse prediction map with local context information for a subtle saliency map. (Zhang et al., 2018) propose a progressive network which consists of multi-path recurrent connections and attention modules. The recurrent structure is used to improve the side-outputs of the backbone network. Then spatial and channel-wise attention mechanisms are used to assign more weights to foreground regions. (A. Islam, 2018) utilizes a framework to integrate three saliency tasks including detection, ranking and subitizing. They consider not only detect salient objects but also predict the total number and the rank order of them by the proposed framework.

In this paper, we use a self-attention saliency network to conduct saliency detection. We utilize a self-attention module, which is calculated by the layer inputs, to enhance the salient semantics of deep layers from layer-level. A refined side-outputs by gated units are used to help network recover the resolution and generate exact saliency maps. The saliency maps can be segmented to the regions with clear boundaries, which can indicate objects locations. Therefore, we can use the saliency maps to enrich the location cues thus improving the performance of the weakly-supervised image segmentation.
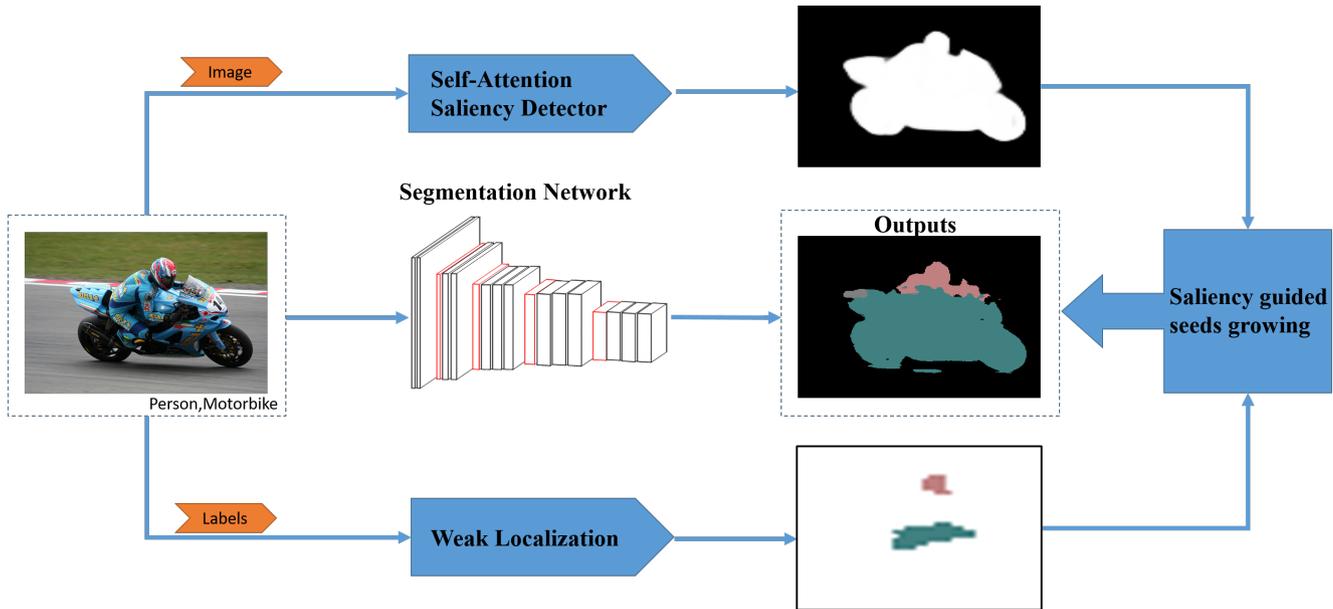
Fig. 2. The overall framework of our network.The saliency guided seeds grown regions are used to train the segmentation network.

2.2. Weakly-supervised image segmentation

Recently, many researches are emerging in weakly-supervised image segmentation (Wu et al., 2018c). These researches achieve significant performance. There are some different kinds of weak labels, such as image labels(Kolesnikov and Lampert, 2016; Huang et al., 2018), points (Bearman et al., 2016), scribbles(Lin et al., 2016). In this section, we mainly introduce the weakly-supervised segmentation from image labels. (Wei et al., 2017a) propose a adversarial erasing approach to locate the object regions. The approach starts with a single small object region, then the region will be erased in an adversarial manner for discovering new and complement object regions. An online learning is also developed to enhance the adversarial erasing approach. (Qi et al., 2016) use objectness information to enhance the performance of weakly-supervised segmentation. One hand, The proposed method use a segmentation network to generate objectness proposals. On the other hand, the proposals are aggregated with object localizations to guide the segmentation network for a better performance. (Joon Oh et al., 2017; Chaudhry et al., 2017) fuse saliency cues into weakly-supervised segmentation using different methods. (Joon Oh et al., 2017) use a existing saliency method to guide the process of training. (Chaudhry et al., 2017) exploit a novel saliency detection method, and combine saliency information with fully attention maps to segment input images.

In this paper, we propose a novel saliency guided weakly-supervised network. Different with (Joon Oh et al., 2017), we do not regard the pixels segmented by saliency method having the same class. First, classification activation maps are used to generate object location cues from image-level labels. These cues will be used as seeds which have high confidence and precision. Then the saliency information are utilized to help these seeds grow to enrich the object location regions. Therefore, we can use the grown regions to train the network for image segmentation.

3. Proposed Method

In this section, we introduce the method proposed in details. First, we use a the classification activation maps of a weak object location network to obtain location cues from image-level label. Then the saliency cues guide the seeds, i.e. the location cues, to grow based on similarity of pixels such that obtaining more object locations. At last, a deep segmentation network is used to learn segment input image using grown object locations.

3.1. Overall structure

In this section, we will illustrate the overall structure of our method. As shown in Figure 2, the entire network has three components. The first component is a self-attention network to generate saliency maps of inputs. The second component is a semantic segmentation network for segmenting the inputs into regions with different labels. The third component is a weak object localization network to generate location cues as seeds. Besides, a small module is used to conduct seeded regions growing under saliency guidance.

When images feed into the network with its categories, the images and category labels are handled by different component. The category labels are transferred to the weak object location network to generate sparse but reliable location cues. The network utilize classification activation maps, which fuse the last convolutional feature maps with their response weights to the images' categories, to extract the location cues. These cues have a high precision but many of them are scattered. To address this problem, we use a saliency guided region growing method to extend the location cues. A saliency detection method based on self-attention are utilized to produce saliency maps which are used to help extend location cues. The saliency network will assign each element in deep layers a self-attention weight to emphasize salient foreground pixel and alleviate the interference of background regions. Then a saliency guided

seed region growing method can be utilized to extend location cues. The growing method not only considers the similarity between pixels but also takes account of their saliency values thus can obtain dense location labels.

Therefore, the results of segmentation network could be supervised by the dense location labels from the seeded region growing method. In the segmentation network, we use a modified VGG16 model pre-trained on ImageNet dataset. The last fully convolutional layer are used to conduct segmentation by a softmax function. For making the boundaries of segmentation more clear, we construct a fully-connected conditional random fields with unary potentials given by the predictions, and pairwise potentials of fixed parametric form which is from input images pixels. In this way, the segmentation network could classify each pixel's category of input images from image-level labels.

3.2. Seeds generation from classification activation maps

We utilize a deep network to detect discriminative object locations as seed cues under image-level labels. Recently, there are many different methods proposed to locate object regions from image-level label, such as multiple instance learning (Pinheiro and Collobert, 2015). Driven by the progress of deep learning, many researches have focused on predict object locations with a deep network. And it can be observed that high-quality seeds, i.e. discriminative object regions, can be obtained by the feature maps of a classification network under the supervision of image-level categories. (Zhou et al., 2016) propose a fully convolutional classification network to predict seed regions using classification activation maps from image category. These activation maps from deep layers generally contain abundant object regions information for robust object localization. Therefore, we employ the seed generation method using classification activation maps.

The input images are feed into a network which is the modified VGG16 network. In this network, the fully connected layers of VGG16 are removed and we use conv7 to represent the last convolutional layers before the final output layer for convenience. The feature maps in conv7 contain abundant location information that are not utilized effectively. A global average pooling (GAP) are used to calculate the spatial average of each feature map in conv7. Then a weighted sum of these GAP values is used to generate the final output, i.e. the image-level category. These weights represent the importance of the GAP values of different feature maps to the image-level category. Therefore, they also can be used to weight the feature maps in conv7 thus helping identify the importance of different image regions to the image category. Then the regions with high importance are used as object location cues, and the weighted feature maps are the classification activation maps.

For a given image, the kth feature map in conv7 can be represented as f_k, then $f_k(x, y)$ denotes the value at location (x, y). The the result F_k of the kth feature map after global average pooling is as following:

$$F_k = \sum_{(x,y)} f_k(x, y) \qquad (1)$$

Thus, for a given category c, the value will be feed into softmax can be represented as:

$$S_c = \sum_k w_k^c F_k \qquad (2)$$

where w_k^c is the weight of F_k for the image category c, indicate the importance of the kth feature map for the image category c. Therefore, the classification activation maps M_c for category c is given by

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \qquad (3)$$

Thus, $M_c(x, y)$ directly indicates the importance of location (x, y) leading to classify the input image to category c.

3.3. Saliency guided seeded region growing

The location cues generated from classification activation maps have a high precision and confidence. There exists a notable problem that these cues are very sparse and small. As reported in Huang et al. (2018), only about 40% pixels in the seeds have labels. Such sparse data can not have a significant improvement in semantic segmentation. Therefore, we want to extend these cues to obtain denser location information. A simple idea is grow the location cues as seeds to unlabeled regions, i.e. seeded region growing method. The seeded region growing method will choose some pixels as initial seeds which are generally selected following by low-level image property, such as color information and texture (Wang and Wu, 2018). Then the method starts from a seed, and seeks the neighborhood to obtain homogeneous image regions by calculating the similarity between seeds and its neighbor pixels. Since the location cues have been generated by classification activation maps, we use these location cues as seeds to obtain denser regions.

Although the seeded region growing method could extend object locations effectively. It may produce error grown pixels because there is lack of constraint condition when seeds growing. For example, the object seeds may grow to a background pixel if the pixel is adjacent to the seeds and has a similar appearance with the seeds. This may cause over-segmentation of discriminative regions. What's more, if background pixels are labeled as object regions, they may grow to adjacent homogeneous regions, i.e. other background pixels. This may affect the quality of segmentation. Therefore, we propose a saliency guided seeded region growing method. Saliency information is an image inherent property and generally presented by saliency maps which indicate the saliency value of each pixel. Salient regions segmented from saliency maps have a clear boundary such that can be used to guide the process of seeds growing.

For a given image I and its corresponding saliency map S. The similarity between two pixels $(x_i, y_i) and (x_j, y_j)$ are defined as following

$$sim_{i,j} = w_{i,j} \left\| I(x_i, y_i) - I(x_j, y_j) \right\| \qquad (4)$$

where $I(x, y)$ represent the pixel value at location (x, y), and $w_{i,j}$ is the saliency weights to control the groing. We use a HSV color space information to calculate the similarity. And the saliency weight $w_{i,j}$ are as following

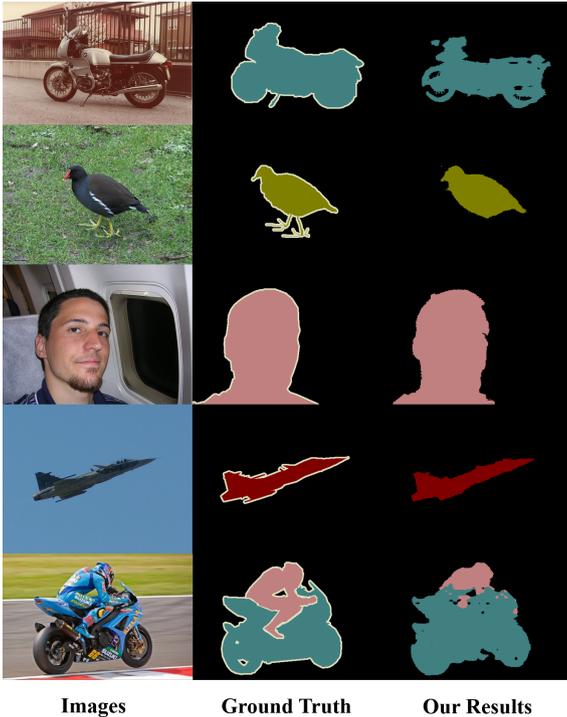$$w_{i,j} = exp(|S(x_i, y_i) - S(x_j, y_j)|) \qquad (5)$$

Images **Ground Truth** **Our Results**

Fig. 3. The results of our methods with ground truths.

where $S(x, y)$ represents the value of the pixel at (x, y) in the saliency map. Then the growing similarity criteria P is given by

$$P_{i,c}(\theta) = \begin{cases} True & if \ sim_{i,c} < \theta \\ False & otherwise \end{cases} \quad (6)$$

where c is the seed pixels with a label. If only the value of $P_{i,c}$ is true and pixel i at location (x_i, y_i) is adjacent to pixel c, the pixel i can be assigned the same label with pixel c. The saliency weight will make the pixels with similar saliency can be broadcast easier such that the grown regions will accord with the shape of the salient objects.

4. Experiments

4.1. Dataset and metrics

We evaluate our model on the PASCAL VOC 2012 image dataset. There are several different tasks benchmark, and we use the segmentation class dataset to demonstrate the effectiveness of our method. The segmentation class dataset has three parts, including training set, validation set and testing set. The training set has 1464 images in total, and the other two sets have 1449 and 1456 images respectively. In a common practice, we augment the training set as the suggestion of Ref. (Hariharan et al., 2011). Therefore, the final training dataset we used in this paper has 10,582 images with weak image-level labels. The validation and testing set are used to evaluate our method with other approaches. For the validation set, the ground truths are available such that we can use to generate the examples of prediction of our method. And for the testing set, the ground truths are not publicly available. Therefore, we submit the results of

Table 1. Comparison of different methods on PASCAL VOC 2012 validation and testing sets (mIoU in %).

Method	Training Images	Val	Test
MIL-FCN (Pathak et al., 2014)	10K	25.7	24.9
CCNN (Pathak et al., 2015)	700K	35.3	35.6
MIL-bb (Pinheiro and Collobert, 2015)	700K	37.8	37.0
EM-Adapt (Papandreou et al., 2015)	10K	38.2	39.6
SN-B (Wei et al., 2016)	10K	41.9	40.6
MIL-seg (Pinheiro and Collobert, 2015)	700K	42.0	43.2
DCSM (Shimoda and Yanai, 2016)	10K	44.1	45.1
BFBP (Saleh et al., 2016)	10K	46.6	48.0
STC (Wei et al., 2017b)	50K	49.8	51.2
Ours	10K	50.5	51.3

testing set to the official PASCAL VOC evaluation server to evaluate the performance of our method.

We adopt the standard intersection over union (IoU) criterion to evaluate a prediction and corresponding ground truth image (Wang et al., 2016). For a given image, P and G present the prediction image and ground truth image respectively. Then, the IoU of the prediction for this image can be given by

$$IoU = \frac{P \cap G}{P \cup G} \quad (7)$$

We use the value of IoU to evaluate the performance in a image. Mean intersection over union (mIoU), which is the average IoU value of a dataset, can be used to evaluate the performance of a method in a dataset.

4.2. Experiment settings

The classification network we used to generate location cues is a slightly modified VGG16 network as the suggestions of Ref. (Kolesnikov and Lampert, 2016). The segmentation network we choose in this paper is the DeepLab-CRF-LargeFOV network which is introduced in Ref. (Chen et al., 2014). The initial weights of these network are pre-trained on the ImageNet dataset (Deng et al., 2009). Seeding losses introduced in Ref. (Kolesnikov and Lampert, 2016) are used to calculate the losses between the segmentation outputs and the grown seeded regions. Stochasitc gradient descent optimizer is used for training the segmentation network with mini-batch. We use the momentum of 0.9 and a weight decay of 0.0005. The size of mini-batch is 4 and the weight decay parameter is 0.0005. We set a dropout rate 0.5 for the last two convolutional layers of segmentation network. The initial learning rate is 1e-3 and it is will be decreased by a factor of 10 every 10 epochs.

In the seed generation, the pixels, whose values in the classification activation maps are in the top 20%, are used as the object location cues. The corresponding saliency maps are generated by Ref. Sun et al. (2018). The parameter θ is the saliency

Table 2. Detailed results of different method on PASCAL VOC 2012 dataset (mIoU in %).

Val set	SFR	EM-adapt	CCNN	MIL-seg	Ours	Test set	CCNN	MIL-seg	RSP	Ours
background	71.7	67.2	68.5	77.2	83.8	background	71	74.7	74	84.7
aeroplane	30.7	29.2	25.5	37.3	59.2	aeroplane	24.2	38.8	33.1	58.5
bicycle	30.5	17.6	18.0	18.4	27.0	bicycle	19.9	19.8	21.7	27.0
bird	26.3	28.6	25.4	25.4	64.3	bird	26.3	27.5	27.7	66.2
boat	20.0	22.2	20.2	28.2	26.4	boat	18.6	21.7	17.7	24.0
bottle	24.2	29.6	36.3	31.9	39.0	bottle	38.1	32.8	38.4	45.7
bus	39.2	47.0	46.8	41.6	67.4	bus	51.7	40.0	55.8	68.8
car	33.7	44.0	47.1	48.1	57.9	car	42.9	50.1	38.3	54.3
cat	50.2	44.2	48.0	50.7	71.8	cat	48.2	47.1	57.9	71.2
chair	17.1	14.6	15.8	12.7	22.6	chair	15.6	7.2	13.6	22.7
cow	29.7	35.1	37.9	45.7	52.5	cow	37.2	44.8	37.4	55.3
diningtable	22.5	24.9	21.0	14.6	24.4	diningtable	18.3	15.8	29.2	22.6
dog	41.3	41.0	44.5	50.9	62.6	dog	43.0	49.4	43.9	66.5
horse	35.7	34.8	34.5	44.1	54.8	horse	38.2	47.3	39.1	59.0
motorbike	43.0	41.6	46.2	39.2	60.8	motorbike	52.2	36.6	52.4	71.4
person	36.0	32.1	40.7	37.9	53.8	person	40.0	36.4	44.4	55.3
pottedpiant	29.0	24.8	30.4	28.3	35.0	pottedpiant	33.8	24.3	30.2	35.2
sheep	34.9	37.4	36.3	44.0	63.6	sheep	36.0	44.5	48.7	58.7
sofa	23.1	24.0	22.2	19.6	31.8	sofa	21.6	21.0	26.4	38.8
train	33.2	38.1	38.8	37.6	47.4	train	33.4	31.5	31.8	39.9
TVmonitor	33.2	31.6	36.9	35.0	51.8	TVmonitor	38.3	41.3	36.3	52.1

guided seeded region growing method is set to 10. And we use the setting of Ref. Krähenbühl and Koltun (2011) to initialize the parameters of conditional random fields (CRFs). The CRFs are used to help generate final outputs of segmentation network, and recover the boundaries information of objects when upscaling the final output segmentations in the testing.

4.3. Comparisons with other methods

We summarize some weakly-supervised image segmentation method, and show their results on PACAL VOC 2012 dataset in Table. 1, including MIL-FCN (Pathak et al., 2014), CCNN (Pathak et al., 2015), MIL-bb (Pinheiro and Collobert, 2015), EM-Adapt (Papandreou et al., 2015), SN-B (Wei et al., 2016), MIL-seg (Pinheiro and Collobert, 2015), DCSM (Shimoda and Yanai, 2016), BFBP (Saleh et al., 2016), STC (Wei et al., 2017b). The mIoU of different methods on validation set and testing set are shown with their training images. The table illustrates that our method has a highest mIoU score of these methods on both validation and testing datasets. We provide these

results for reference and indicate the number of training images they used. Some methods are trained on different training sets or with different kinds of annotations, such as bounding boxes and image-level labels. Among the approaches, CCNN, MIL-bb and MIL-seg use a larger training set including 700K images. Mil-seg and SN-B implicitly utilize pixel-level supervision in the training phase.

Table 2 shows the detailed results. The mIoU values of each category on validation and testing datasets demonstrate the effectiveness of our method. We compare our method with some methods including SFR (Kim and Hwang, 2016), RSP (Krapac and Šegvić, 2016), CCNN, MIL-seg. The mIoU values of our method are the highest in most categories.

4.4. Qualitative results

Figure 3 shows some successful segmentation results. It shows our method can produce accurate segmentations even for complicated images and recover ne details of the boundary. It can be observed that the results of out method is very close to

the ground truths. In the first four rows, we use four single objects to illustrate the effectiveness of saliency guidance. In the bottom row, there are two categories in the image, our method also can generate a satisfactory segmentation.

5. Conclusion

In this paper, we propose a novel method to segment images from image-level labels. The object location cues are generated by a localization network using the image category. Then a saliency guided seeded region growing method is used to extend these location cues. Therefore, the grown regions can be used to train a segmentation network for a better performance.

Acknowledgments

References

A. Islam, M. Kalash, N.D.B.B., 2018. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects, in: Computer Vision and Pattern Recognition (CVPR).

Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L., 2016. What's the point: Semantic segmentation with point supervision, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 549–565.

Chaudhry, A., Dokania, P.K., Torr, P.H.S., 2017. Discovering class-specific pixels for weakly-supervised semantic segmentation. CoRR abs/1707.05821. URL: arxiv.org/abs/1707.05821.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. CoRR abs/1412.7062. URL: http://arxiv.org/abs/1412.7062.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Hariharan, B., Arbelez, P., Bourdev, L., Maji, S., Malik, J., 2011. Semantic contours from inverse detectors, in: 2011 International Conference on Computer Vision, pp. 991–998. doi:10.1109/ICCV.2011.6126343.

Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J., 2018. Weakly-supervised semantic segmentation network with deep seeded region growing, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Joon Oh, S., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B., 2017. Exploiting saliency for object segmentation from image level labels, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kim, H.E., Hwang, S., 2016. Deconvolutional Feature Stacking for Weakly-Supervised Semantic Segmentation. ArXiv e-prints .

Kolesnikov, A., Lampert, C.H., 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 695–711.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials, in: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp. 109–117.

Krapac, J., Šegvić, S., 2016. Weakly-supervised semantic segmentation by redistributing region scores back to the pixels, in: Rosenhahn, B., Andres, B. (Eds.), Pattern Recognition, Springer International Publishing, Cham. pp. 377–388.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3159–3167. doi:10.1109/CVPR.2016.344.

Papandreou, G., Chen, L., Murphy, K., Yuille, A.L., 2015. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. CoRR abs/1502.02734. URL: http://arxiv.org/abs/1502.02734.

Pathak, D., Krhenbhl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1796–1804. doi:10.1109/ICCV.2015.209.

Pathak, D., Shelhamer, E., Long, J., Darrell, T., 2014. Fully convolutional multi-class multiple instance learning. CoRR abs/1412.7144. URL: http://arxiv.org/abs/1412.7144.

Pinheiro, P.O., Collobert, R., 2015. From image-level to pixel-level labeling with convolutional networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1713–1721. doi:10.1109/CVPR.2015.7298780.

Qi, X., Liu, Z., Shi, J., Zhao, H., Jia, Jiaya", e.B., Matas, J., Sebe, N., Welling, M., 2016. Augmented feedback in semantic segmentation under image level supervision, in: ECCV 2016, Springer International Publishing, Cham. pp. 90–105.

Saleh, F., Aliakbarian, M.S., Salzmann, M., Petersson, L., Gould, S., Alvarez, J.M., 2016. Built-in foreground/background prior for weakly-supervised semantic segmentation, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 413–432.

Shimoda, W., Yanai, K., 2016. Distinct class-specific saliency maps for weakly supervised semantic segmentation, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 218–234.

Sun, F., Li, W., Guan, Y., 2018. Self-attention recurrent network for saliency detection. Multimedia Tools and Applications URL: https://doi.org/10.1007/s11042-018-6591-3, doi:10.1007/s11042-018-6591-3.

Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H., 2017a. A stagewise refinement model for detecting salient objects in images, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4039–4048. doi:10.1109/ICCV.2017.433.

Wang, Y., Lin, X., Wu, L., Zhang., W., 2017b. Effective multi-query expansions: Collborative deep networks for robust landmark retrieval. IEEE Transactions on Image Processing 26, 1393–1404.

Wang, Y., Lin, X., Wu, L., Zhang, W., Zhang, Q., Huang, X., 2015. Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Transactions on Image Processing 24, 3939–3949.

Wang, Y., Wu, L., 2018. Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. Neural Networks 103, 1–8.

Wang, Y., Wu, L., Lin, X., Gao, J., 2018. Multiview spectral clustering via structured low-rank matrix factorization. IEEE Transactions on Neural Networks and Learning Systems 29, 4833–4843.

Wang, Y., Zhang, W., Wu, L., Lin, X., Fang, M., Pan, S., 2016. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering, in: IJCAI 2016, pp. 2153–2159.

Wei, Y., Feng, J., Liang, X., Cheng, M., Zhao, Y., Yan, S., 2017a. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. CoRR abs/1703.08448. URL: http://arxiv.org/abs/1703.08448.

Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y., Yan, S., 2016. Learning to segment with image-level annotations. Pattern Recognition 59, 234–244. doi:https://doi.org/10.1016/j.patcog.2016.01.015. compositional Models and Structured Learning for Visual Recognition.

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M., Feng, J., Zhao, Y., Yan, S., 2017b. Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2314–2320. doi:10.1109/TPAMI.2016.2636150.

Wu, L., Wang, Y., 2018. What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. Pattern Recognition 76, 727–738.

Wu, L., Wang, Y., Gao, J., Li, X., 2018a. Deep adaptive feature embedding with local sample distributions for person re-identification. Pattern Recognition 73, 275–288.

Wu, L., Wang, Y., Gao, J., Li, X., 2018b. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Trans. Multimedia .

Wu, L., Wang, Y., Li, X., Gao, J., 2018c. Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE Transactions on

Cybermetics .

Wu, L., Wang, Y., Shao, L., 2018d. Cycle-consistent deep generative hashing for cross-modal retrieval. CoRR abs/1804.11013. URL: `http://arxiv.org/abs/1804.11013`.

Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G., 2018. Progressive attention guided recurrent network for salient object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929. doi:`10.1109/CVPR.2016.319`.