Weakly Supervised Semantic Segmentation with Patch-Based Metric Learning Enhancement

Patrick P. K. Chan[1(✉)], Keke Chen[1], Linyi Xu[1], Xiaoman Hu[1], and Daniel S. Yeung[2]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
patrickchan@scut.edu.cn
[2] Hong Kong, China

Abstract. Weakly supervised semantic segmentation (WSSS) methods are more flexible and less costly than supervised ones since no pixel-level annotation is required. Class activation maps (CAMs) are commonly used in existing WSSS methods with image-level annotations to identify seed localization cues. However, as CAMs are obtained from a classification network that mainly focuses on the most discriminative parts of an object, less discriminative parts may be ignored and not identified. This study aims to improve the local visual understanding on objects of the classification network by considering an additional metric learning task on patches sampled from each CAM-based object proposal. As the patches contain different object parts and surrounding backgrounds, not only the most discriminative object parts but the entire objects are learned through leveraging the patch similarity. After the joint training process with the proposed patch-based metric learning and classification tasks, we expect more discriminative local features can be learned by the backbone network. As a result, more complete class-specific regions of an object can be identified. Extensive experiments on the PASCAL VOC 2012 dataset validate the superiority of our method. Our proposed model achieves improvement compared with the state-of-the-art methods.

Keywords: Semantic segmentation · Fully convolutional neural networks · Weakly supervised learning · Metric learning

1 Introduction

The quality and quantity of labeled samples play an essential role in achieving satisfactory performance for deep learning. However, obtaining sufficient training samples with labels is expensive and time-consuming, especially for supervised semantic segmentation since pixel-level annotation is required. In contrast, WSSS requires only lower-level annotations, like bounding boxes [1], scribbles [2], and image-level labels [3]. Compared to other kinds of annotation, image-level annotation is more economical and easily to be obtained since it is available in many open large-scale image datasets, *e.g.*ImageNet [4].

I. Farkaš et al. (Eds.): ICANN 2021, LNCS 12893, pp. 471–482, 2021.
https://doi.org/10.1007/978-3-030-86365-4_38

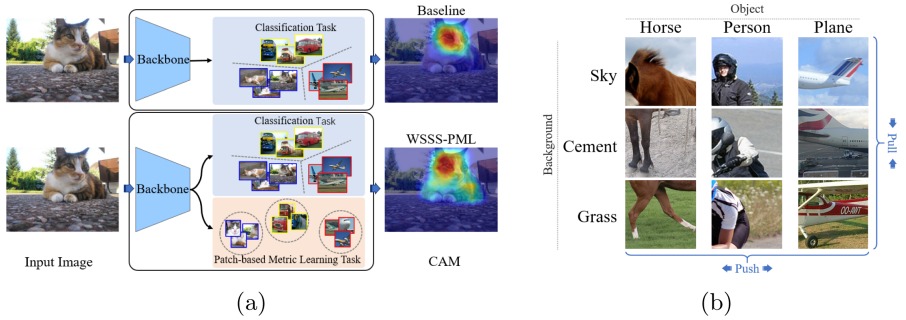

Fig. 1. (a) Examples of CAMs generated by the baseline (only relying on the classification task) and our proposed WSSS-PML (relying on both the classification and patch-based metric learning tasks). (b) Examples of the patches sampled from CAM-based object proposals. The proposed patch-based metric learning task pulls patch pairs with the same objects closer and pushes patch pairs with different objects away. More class-specific object regions can be learnt.

Most of the advanced WSSS methods with image-level annotations follow the pipeline with three steps. The localization cues for each object class are identified by the response maps, *e.g.*Classification Activation Map (CAM) [5], inferred from a multi-label classification network trained with image-level labels. The localization cues, which are treated as the seed areas, are then expanded and refined as the Pseudo Masks. Finally, a standard fully-supervised semantic segmentation model is trained by using the Pseudo Masks as the ground truth and is used in the inference stage. The first step of the three-step pipeline is the most important since it serves as supervision. The CAM is a prevailing method to obtain seed localization cues identified according to the contribution of a pixel to the decision of the classification network. However, directly relying on the classification network may cause incomplete and sparse seed areas. Since an image can be classified correctly by only focusing on most discriminative parts, the contribution of less discriminative regions may be similar to the surrounding background in CAM, *i.e.*less discriminative regions cannot be identified precisely in the seed areas. Figure 1(a) illustrates an example of a cat image in the PASCAL VOC 2012 dataset. If only the classification task is considered, the activation scores on the cat's body are small as the surrounding background since its head provides necessary discriminant information.

A patch-based metric learning task for weakly supervised semantic segmentation (WSSS-PML) which considers the standard classification task and also patch-based metric learning task is devised. The similarity of the patches sampled from the CAM-based object proposals is learnt in the metric learning task in which the same-category patches are pulled together while the different-category patches are pushed away in the latent feature space. As some patches contain different objects and similar backgrounds, shown in Fig. 1(b), this mechanism forces the backbone network to focus on local patterns of the target objects and

ignore the backgrounds. As both global and local features are learnt by the classification task and patch-based metric learning task respectively, it is expected that object regions should be explored more thoroughly in our model.

The shared backbone network of the proposed WSSS-PML is first learnt from the classification task to obtain reliable CAM-based object proposals. Patches are selected with non-overlapping sliding window to completely cover the object proposals generated by the trained backbone network. The hard sampling strategy [6], which pushes the closest negative patch pairs and pulls the furthest positive patch pairs to form a denser response to the target object, is applied to form triplets in each mini-batch. The backbone network is then updated by both classification and patch-based metric learning tasks. The framework of our proposed model is shown in Fig. 2. Extensive experiments have been conducted on the PASCAL VOC 2012 dataset to demonstrate the effectiveness of our WSSS-PML. The results confirm that the high quality of seed localization cues provided by our method. Furthermore, overall semantic segmentation is better by using generated seed localization cues compared to state-of-the-arts.

The main contributions of this paper are summarized as follows:

- A novel patch-based metric learning task in addition to the classification task is proposed to enhance the visual understanding on local object parts from the sampled patches to identify more complete seed areas.
- The patch-based metric learning task forces the backbone network to learn on object parts by minimizing the triplet loss with the hard sampling strategy. It is expected more class-specific object regions can be identified.
- The experimental results on PASCAL VOC 2012 dataset demonstrate the outstanding performance of our method on improving the seed area quality, and the performance boost is achieved compared to state-of-the-art methods.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

Most of WSSS methods with image-level annotations follow a three-step pipeline. The studies of the first two steps, the seed areas identification and refinement methods for generating Pseudo Masks, are introduced.

Seed Areas Identification: CAM [5] is widely used in seed areas identification stage. The contribution of each pixel to the image classification task is quantified to localize the object regions. Several methods aiming at generating high-quality seed areas have been proposed and can be categorized into two types. The first type of the methods pays extra attention on the less discriminative regions by ignoring the knowledge of the network [7,8]. However, the object parts may not be targeted since the dropout units are chosen randomly, *i.e.*there is no guarantee that less discriminative regions of an object can be learnt better. Another kind considers additional constraints in the training of the classification network. Siamese network (SSENet) [9] forces the same CAM should be output on the

images with different spatial transformation due to its equivariance. SEAM [10] extends SSENet by considering a pixel correlation module (PCM) additionally to refine CAM. Sub-Category Exploration [11] (SC-CAM) exploits the sub-category relations between object classes through iteratively clustering on image features. More object parts are expected to be indicated by applying a more challenging sub-category classification task. However, an unstable clustering process may require additional iterations for training in SC-CAM. Different from the methods mentioned above, our method enforces the backbone network to identify local object parts from the sampled patches by using metric learning.

Seed Areas Refinement: Seed areas refinement aims to expand and constraint the coarse response map to cover whole object areas precisely. AffinityNet [12], one of the most common refinement methods recently, learns the semantic affinity between adjacent pixel pairs and propagates local activation scores to an entire object through random walk. Another pixel-level affinity method, IRNet [13], aims at detecting accurate object boundaries and propagating activation scores within the boundaries. DSRG [14] reveals more complete areas by integrating the seed region growing algorithm with a deep segmentation network to expand the seed areas iteratively. Moreover, the cross-image relationship is investigate by CIAN [15] and ICD [16], which propose end-to-end framework to indicate more class-specific pixels. Context Adjustment (CONTA) [17] is an external mechanism for seed areas refinement which reconsiders the seed areas after the pipeline is completed by removing the confounding bias of context in classification.

2.2 Metric Learning

Metric Learning has been widely used in computer vision tasks, *e.g.*landmark recognition, face recognition and person re-identification [18]. It aims to learn an embedding space in which samples belong to the same class are close together and those of different classes are far apart. The triplet loss [19], which is a popular objective function used in Metric Learning, minimizes intra-class variations and maximizes inter-class variations simultaneously by using triplets generated by positive and negative samples of an anchor. In order to enhance the training efficiency, a hard sampling [6] is proposed to choose the farthest positive sample and nearest negative sample in a triplet.

3 Weakly Supervised Semantic Segmentation with Patch-Based Metric Learning Enhancement

Our proposed WSSS-PML focuses on seed areas identification in the WSSS pipeline. The backbone network is first trained on a standard classification task to generate reasonable object proposals (Sect. 3.1). CAM of each class is calculated from the trained backbone network. The object proposals are identified according to the seed localization cues calculated by using the CAMs

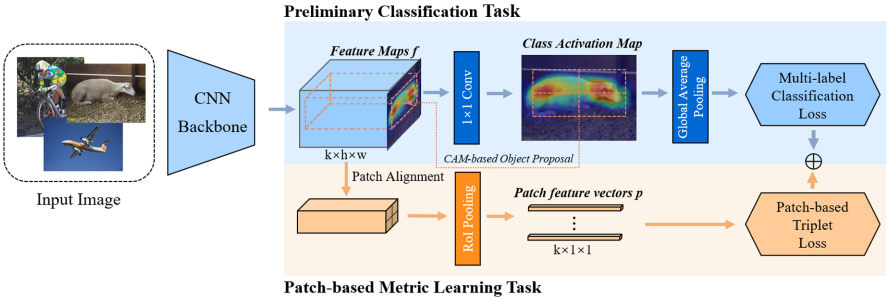

Fig. 2. Framework of our proposed model. The network is first initialized by minimizing the classification loss. Both the proposed triplet loss of the metric learning on the patches and the classification loss are then optimized.

(Sect. 3.2). To enhance the learning on less discriminative regions, patches are sampled from each CAM-based object proposal for the patch-based metric learning task (Sect. 3.3). The backbone network is adjusted according to the triplet loss ($Loss^{p}_{tri}$) using the hard sampling strategy [6] jointly with the loss ($Loss_{cls}$) of the standard classification task (Sect. 3.4). The trained backbone network is used in inference to identify seed areas. The framework of our model is shown in Fig. 2.

3.1 Classification Task

A convolutional neural network containing a global average pooling and fully-connected classification layers is used as the classification network. Given an image I, the feature map output by the backbone network is denoted as $f = \phi(I) \in \mathbb{R}^{k\times h\times w}$, where ϕ denotes the function of the backbone network, f denotes the output feature map, k, h and w represent the channels, height and width of the feature map respectively.

The backbone network is updated by minimizing the multi-label classification loss $Loss_{cls}$ of the preliminary classification task. The CAM of object class c can be obtained according to the trained classification network as follows:

$$M_c(x, y) = \theta_c^{\top} f(x, y) \tag{1}$$

where (x, y) denotes the coordinate on f, θ_c indicates the classification weight of c, and $^{\top}$ represents the inner product of two vectors. $M_c(x, y)$ which represents the activation score of (x, y) belonging to c is normalized linearly to $[0, 1]$. CAMs are then upsampled to meet with the height H and width W of I by the bilinear interpolation, defined as $M \in \mathbb{R}^{C\times H\times W}$, where C is the object class number.

3.2 CAM-Based Object Proposal

As the training data of the patch-based metric learning task, the object proposals are extracted from CAMs during the forward pass of the network in each mini-batch. To extract the object proposals, the CAM is first transformed to the Class

Label Map (CLM), denoted by $\mathbf{L} \in \mathbb{R}^{H \times W}$, shown in Eq. (2). Each coordinate (x, y) is assigned to $c \in C$ according to the class with the maximum activation score which is larger than a pre-selected threshold β, otherwise, the pixel is classified as the background class denoted by 0. CLM is formulated as:

$$\mathbf{L}(x,y) = \begin{cases} \arg\max\limits_{1 \le c \le C} M_c(x,y), & \max\limits_{1 \le c \le C} M_c(x,y) > \beta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{L}(x, y)$ denotes the class label of the coordinate (x, y) on the CAM, while $M_c(x, y)$ denotes the activation score that the coordinate (x, y) belongs to the class c. β serves as a hard threshold to separate the background and foreground objects, *i.e.*a coordinate with an activation score smaller than β is classified as the background; otherwise, it is classified as a foreground object.

The quality of the object proposal from which the patches are sampled in our model is affected by β significantly. If β is small, some over-activated background areas will be identified wrongly as a foreground object. On the other hand, non-discriminative object areas may be excluded when β is large. The minimum bounding box containing the connected pixels with the same class label is defined as an object proposal. As an image may contain more than one bounding box for a class, Non-Maximum Suppression (NMS) [20] is adopted to remove proposals of the same class with a large overlapping rate. Moreover, an object proposal with a small size or large aspect ratio is filtered in order to reduce the noise influence.

3.3 Patch-Based Metric Learning

p patches are sampled from each CAM-based object proposal with a non-overlapping sliding window to cover the entire object proposal. A sampled patch is aligned back to the feature map f as the Region of Interest (RoI). As a result, a k-dimensional feature vector $P \in \mathbb{R}^k$ is obtained from f via the RoI pooling for each patch. The label of a patch is then set as the class of its object proposal.

Hard Sampling Strategy in Mini-Batch. Let $\mathcal{P} = \{P^i\}_{i=1}^M$ be the set of patch feature vectors in a mini-batch, where M is the number of the sampled patches. The hard sampling strategy [19] is applied to select the triplet from each $\mathcal{P}$. Unlike the random sampling method, the furthest patch of the same class and the closest patch of a different class are chosen to form a triplet in the hard sampling method in order to obtain a more condensed feature space. The training objective of the patch-based metric learning task is shown in Eq. (3).

$$Loss_{tri}^{p} = \frac{1}{M} \sum_{i=1}^{M} \max\{0, D(P_a^i, P_+^i) - D(P_a^i, P_-^i) + \alpha\} \tag{3}$$

where P_+^i and P_-^i denote the positive and negative pathes of the anchor patch P_i^a, α represents the margin, D denotes the distance measure, and the Euclidean distance is used in our study.

3.4 Joint Training Scheme

The losses of the proposed patch-based metric learning and standard classification tasks are minimized simultaneously in our model, shown as follows:

$$Loss = Loss_{cls} + \lambda Loss^{p}_{tri} \tag{4}$$

where λ denotes the tradeoff between two losses and aims to balance them.

4 Experiment

4.1 Experimental Setting

Dataset: The PASCAL VOC 2012 segmentation benchmark [21] containing 20 object and 1 background classes is considered in our study. 1,464, 1,449 and 1,456 images are contained in training, validation and test sets respectively. The augmentation training set which includes 10,582 images from SBD [22] is applied to enlarge the training set as the common practice. Only image-level labels are used during training.

Evaluation: Mean intersection over union (mIoU), recall and precision are for evaluation following previous works.

Implementation Details: The revised ResNet38 pre-trained on ImageNet is used as the backbone network of our model by following AffinityNet [12]. The data augmentation strategy of [12] is considered in the training process. The default parameters of [12] are applied to train the preliminary classification task. The multi-label soft margin loss is adopted as $Loss_{cls}$ for the classification task in our model. The learning rate r is set to 0.1 with 0.0005 as weight decay in Adam optimizer in the first training on the classification task. During the joint training of the patch-based metric learning and the classification tasks, the learning rate is set to 0.01 and halved every epoch with adam optimizer. The number of patches p is set to 4 in the triplet loss $Loss^{p}_{tri}$. λ in Eq. (4) is set as 0.05 to balance $Loss_{cls}$ and $Loss^{p}_{tri}$. β in Eq. (2) is set as 0.2.

Reproducibility: All experiments are implemented with PyTorch and carried out on 2 NVIDIA GTX 1080 Ti GPUs.

4.2 CAM Analysis

CAMs generated by WSSS-PML and the baseline which only considers classification task are compared in this section in terms of mIoU, recall, precision and visualization. It should be noted that in Epoch#0 of WSSS-PML, the backbone network is trained by the classification task which is identical to the baseline.

As illustrated in Table 1, our WSSS-PML outperforms the baseline (Epoch#0) significantly. With only one epoch of joint optimization in Epoch#1, the mIoU, recall and precision of CAMs generated by our method increases 2.6%, 3.3% and 1.5% respectively. The improvement of our method are mainly caused

Table 1. Quality of CAM generated by WSSS-PML with different epoches on PASCAL VOC 2012 training set

Epoch	mIoU (%)	Recall (%)	Precision (%)
Epoch#0 (*baseline*)	47.7	67.4	61.4
Epoch#1	$50.3_{+2.6}$	$70.7_{+3.3}$	$\mathbf{62.9}_{+1.5}$
Epoch#2	$\mathbf{50.8}_{+3.1}$	$72.6_{+5.2}$	$62.3_{+0.9}$
Epoch#3	$\mathbf{50.8}_{+3.1}$	$\mathbf{72.7}_{+5.3}$	$62.3_{+0.9}$

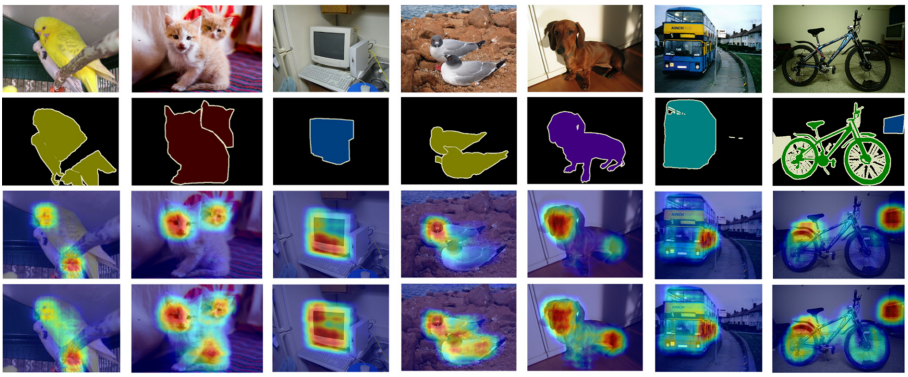

Fig. 3. Visualization examples of CAMs on PASCAL VOC 2012 training set. From the top to bottom, the rows represent the original images, the ground truth, the CAMs generated by the baseline and the CAMs generated by our WSSS-PML respectively.

by the increase of the recall, which means class-specific object parts ignored by the standard classification task are learnt by our WSSS-PML. The performance of our method becomes stable on Epoch#3. For the rest of the experiments, Epoches#2 is adopted considering both the computation efficiency and the performance.

The visualization examples of the generated CAMs on three training images are shown in Fig. 3. The CAM is represented with a heatmap in which the color closer to red means the higher activation score. The results indicate that the baseline tends to activate on the most discriminative regions. Through learning from both classification and patch-based metric learning tasks, the CAM generated by WSSS-PML is capable to cover more complete object regions than the baseline which only considers the classification task.

4.3 State-of-the-Art Methods Comparison

To further evaluate the effectiveness of our model, the AffinityNet [12] is adopted as the seed areas refinement method to generate the Pseudo Masks. One should be noted that WSSS-PML only focuses on the seed areas identification stage. Therefore, any advanced seed areas refinement framework [14–17] could be

Table 2. Comparison of state-of-the-art WSSS seed areas identification methods. 'S' and 'I' denote the saliency map and image level label, while 'val' and 'test' denote the validation and test sets of PASCAL VOC 2012 dataset.

Methods	Supervision	Backbone	mIoU (val)	mIoU (test)
AffinityNet (baseline) [12]	I	ResNet-38	61.7%	63.7%
AffinityNet (baseline)* [12]	I	ResNet-101	**61.9%**	**62.3%**
FickleNet [8]	I+S	ResNet-101	64.9%	65.3%
OAA [23]	I	ResNet-101	63.9%	65.6%
SEAM [10]	I	ResNet-38	64.5%	65.7%
SEAM* [10]	I	ResNet-101	64.4%	64.9%
SC-CAM [11]	I	ResNet-101	66.1%	65.9%
WSSS-PML	I	ResNet-101	**66.5%**	**66.5%**

∗ Our re-implemented results with DeepLab-v2 (ResNet-101) as the segmentation network.

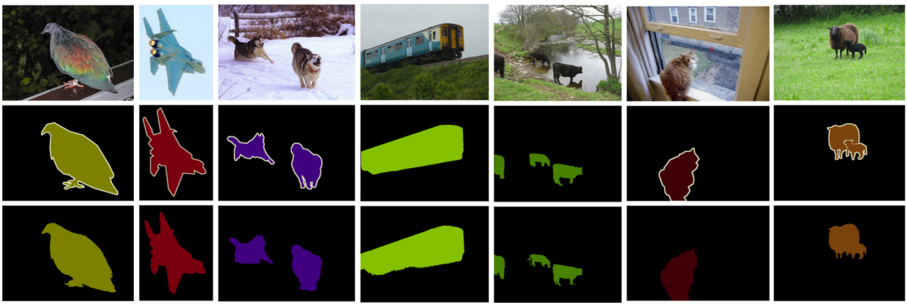

Fig. 4. Visualization results of segmentation mask on PASCAL VOC 2012 validation set. From the top to bottom, the rows represent the original images, the ground truth and the final segmentation mask of our method respectively.

embedded and is not considered in this comparison. For a fair comparison, we adopt the classical DeepLab-v2 with ResNet-101 as the backbone. Some examples of the final semantic segmentation results are illustrated in Fig. 4, which verifies that our results are close to the ground truth segmentation masks.

State-of-the-art seed areas identification methods of WSSS using the ResNet-based backbone are compared with WSSS-PML in Table 2. Our method achieves the highest mIoU on the validation and test sets against all the advanced methods including those using saliency maps as additional supervision. The baseline, *i.e.*AffinityNet, is 4.6% and 4.2% lower than WSSS-PML on validation and test sets respectively, which suggests that the quality of the identified seed areas improves significantly. Table 3 compares the category performance on the validation set of WSSS-PML with AffinityNet [12], SEAM [10], and SC-CAM [11], which adopted the same seed areas refinement method [12]. Our method performs better on 16 out of the 20 categories comparing to the baseline, *i.e.*AffinityNet,

Table 3. Category performance comparisons with advanced seed areas identification methods that adopted AffinityNet as the seed areas refinement method on PASCAL VOC 2012 validation set.

Score	Areo	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
AffinityNet	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
SEAM	68.5	33.3	85.7	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1	79.9	73.8	71.4	75.2	48.9	79.8	40.9	58.2	53.0	64.5
SC-CAM	51.6	30.3	82.9	53.0	75.8	88.6	74.8	86.6	32.4	79.9	53.8	82.3	78.5	70.4	71.2	40.2	78.3	42.9	66.8	58.8	66.1
WSSS_PML	53.5	32.0	84.4	51.8	69.9	88.9	77.9	88.4	34.1	82.5	53.2	84.0	81.4	73.3	71.1	40.7	79.1	44.4	61.3	56.0	66.5

Table 4. CAMs Quality of the baseline, WSSS-PML$_{Ran}$, and WSSS-PML.

Model	PML	Sampling	mIoU (%)
Baseline	–	–	47.7
WSSS-PML$_{Ran}$	√	Random	49.6
WSSS-PML	√	Hard	**50.8**

which demonstrates the effectiveness of our model on different object classes. Our model also achieves the highest and second highest times, at 9 and 6 respectively among all the methods.

All the results confirm that our proposed patch-based metric learning task in the multi-task framework improves the understanding of the backbone network on each part of objects.

4.4 Ablation Study on Hard Sampling Strategy

We analyse the effect of the hard sampling strategy in the proposed patch-based metric learning on the quality of generated CAM. In addition to the baseline, WSSS-PML$_{Ran}$ representing the random sampling in triplet generation of our method is also compared with WSSS-PML using the hard sampling. Table 4 illustrate the mIoU of the generated CAMs. WSSS-PML$_{Ran}$ achieves 1.9% mIoU higher than the baseline, which indicates that the patch-based metric learning enhances the complete understanding on objects of the backbone network. The hard sampling strategy also plays an important role in learning since mIoU of WSSS-PML$_{Ran}$ improves from 49.6% to 50.8% when the hard sampling strategy is used.

4.5 Computation Complexity

Compared to the baseline, WSSS-PML only requires two more epoches for joint training on the classification and patch-based metric learning tasks. the average training time on each image of WSSS-PML is 0.629 sec/img, which is similar to the baseline (0.628 sec/img). It suggests that the patch-based metric learning does not significantly increase the training time complexity. Besides, the inference computation complexity of our approach is the same as the baseline since the backbone networks are the same.

5 Conclusion

One major problem of seed areas identification methods in the WSSS pipeline is that the less discriminative regions of an object may be neglected due to their weak contribution to the classification task. In this paper, we propose a patch-based metric learning task integrated with the classification task for revealing a complete seed areas in WSSS. Our method aims to strengthen the visual understanding on different regions of an object by forcing the backbone network to distinguish object's parts represented by patches of the CAM-based object proposals. With low computation cost, the objects are identified more completely by our proposed method compared with the baseline classification task in the experiments. By using the AffinityNet as seed areas refinement method and DeepLab-v2 as the segmentation model, our method achieves the better performance than the state-of-the-art seed areas identification methods of WSSS on PASCAL VOC 2012 dataset. The experimental results confirm that WSSS-PML improves the understanding of the backbone network on the objects successfully.

Acknowledgments. This paper is supported by the Natural Science Foundation of Guangdong Province, China (No. 2018A030313203).

References

1. Dai, J.F., He, K.M., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1635–1643. IEEE (2015)
2. Lin, D., Dai, J.F., et al.: ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3159–3167. IEEE (2017)
3. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 695–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_42
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Zhou, B.L., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929. IEEE (2016)
6. Wu, C., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2859–2867. IEEE (2017)
7. Wei, Y.C., Feng, J.S., Liang, X.D., Cheng, M.M., Zhao, Y., Yan, S.C: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6488–6496. IEEE (2017)
8. Lee, J., Kim, E., et al.: FickleNet: weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5267–5276. IEEE (2019)

9. Wang, Y.D., Zhang, J., Kan, M.N., Shan, S.G., Chen, X.L.: Self-supervised scale equivariant network for weakly supervised semantic segmentation. arXiv: Computer Vision and Pattern Recognition (2019)
10. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12272–12281. IEEE (2020)
11. Chang, Y.T., Wang, Q.S., Hung, W.C., Robinson, P.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the International Conference on Computer Vision, IEEE (2020)
12. Ahn, J., Kwak, S.: 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, pp. 4981–4990. IEEE (2018)
13. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2204–2213. IEEE (2019)
14. Huang, Z.L., Wang, X.G., Wang, J.S., Liu, W.Y., Wang, J.D.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7014–7023. IEEE (2018)
15. Fan, J.S., Zhang, Z.X., Tan, T.N., Song, C.F., Xiao, J.: CIAN: cross-image affinity net for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10762–10769 (2020)
16. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4282–4291. IEEE (2020)
17. Zhang, D., Zhang, H.W., Tang, J.H., Hua, X.S., Sun, Q.R.: Causal intervention for weakly-supervised semantic segmentation. In: Proceedings of the Conference on Neural Information Processing Systems (2020)
18. Boiarov, A., Tyantov, E.: Large scale landmark recognition via deep metric learning. In: Proceedings of the ACM International Conference, pp. 169–178 (2019)
19. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823. IEEE (2015)
20. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: Proceedings of the International Conference on Pattern Recognition, pp. 850–855. IEEE (2006)
21. Everingham, M., Eslami, S.M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision **111**(1), 98–136 (2015)
22. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
23. Jiang, P.T., Hou, Q.B., Cao, Y., Cheng, M.M., Wei, Y.C., Xiong, H.K.: Integral object mining via online attention accumulation. In: Proceedings of the International Conference on Computer Vision. IEEE (2019)