Person Part Segmentation based on Weak Supervision

Yalong Jiang[1]
yalong.jiang@connect.polyu.hk
Zheru Chi[1]
chi.zheru@polyu.edu.hk

[1]Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, HK

Abstract

In this paper we address the task of semantic segmentation and propose a weakly supervised scheme for person part segmentation. Previous work on segmentation suffers from the lack in supervision due to the expensive labelling process. To address this issue, the correlation between depth estimation and semantic segmentation is explored in this paper and a scheme is proposed to utilize the knowledge learned from depth labels to serve the segmentation task. The supervision in the form of depth maps which can be acquired automatically contributes to the training of the segmentation model. The combination of the two types of supervision allows us to augment the training data of person part segmentation without additional annotations. Quantitative results have shown that the proposed scheme outperforms existing methods by over 5%. Qualitative results have also shown that the proposed scheme significantly outperforms existing methods.

Introduction

Deep Convolutional Neural Networks (CNNs) have brought significant improvements in semantic segmentation [1, 2, 3] where one categorical label is assigned to each pixel in an image. Existing work on segmentation involves Path Aggregation (PA) [4], Large Kernel Matters (LKM) [5], Mask RCNN (MRCNN) [6] and Deeplab-V2 with spatial pyramid pooling [7]. However, the robustness of these methods heavily relies on the supervision provided by training data, and such supervision is expensive to collect [8].

Existing datasets for the semantic segmentation on person parts [9] involves annotations of less than 10 classes and no more than 4,000 images for training and validation, the data is far from enough to train complex CNNs [10, 11] with over 100 layers. What's worse, data augmentation is unrealistic because it is addressed in [8] that labeling an image pixel-by-pixel takes 239.7 seconds on average. However, there are other types of cheap labeling. For instance, the labeling process of an image in classification tasks requires less than 20 seconds. The annotation of depth maps from RGB image pairs with overlapping viewpoints can even be conducted automatically [12]. This motivates the exploration of weakly supervised semantic segmentation in which the supervision from other types of labels serves the segmentation task. Image-level annotations and bounding boxes have been used for improving segmentation [13, 14, 15]. Additionally, scribbles and points have been introduced in [8, 16] as weak supervision.

The supervision from image-level labels or bounding boxes can be regarded as the knowledge that is transferred to segmentation models. Different from other weakly supervised methods, this paper focuses on transferring the knowledge learned from depth

estimation to the segmentation model. To our knowledge, this paper is the first to explore the correlation between dense depth estimation and semantic segmentation and conduct segmentation with the hierarchical features provided by predicted depth maps.

The advantage of utilizing the knowledge learned from depth estimation comes in two ways. Firstly, the amount of data available is huge. Any pair of images with overlapping viewpoints can be processed with multi-view stereo (MVS) to automatically produce dense depth maps. Secondly, depth estimation and semantic segmentation are closely related. The former assigns continuous depth values to pixels while the latter assigns discrete categorical labels to pixels. The predicted depth maps facilitate hierarchical descriptions of images, the hierarchical descriptions are helpful for segmentation.

In this paper we describe a process in which the knowledge learned from depth labels is transferred to the segmentation model. The model for depth estimation is trained first and provides pixel-level predictions which are used as initializations for the segmentation model. For simplicity, we denote Model-Depth and Model-Seg to represent the two models. The key insight is based on the fact that the large-scale training data for depth estimation guarantees good initializations.

The learning process is divided into two steps: (1) Train Model-Depth on the large-scale MegaDepth Dataset [17] which includes Internet photos for depth estimation. The ground truth maps in the dataset are generated automatically with structure-from-motion (SfM). (2) Base both the training and testing of Model-Seg on the predictions from Mode-Depth. The way of transferring knowledge is implemented by adding the predicted depth channel to the input channels of Model-Seg [2, 9]. As is shown in Figure 1, the knowledge learned from the domain of depth estimation is transferred to the domain of segmentation. The input images to Model-Seg are with four channels.

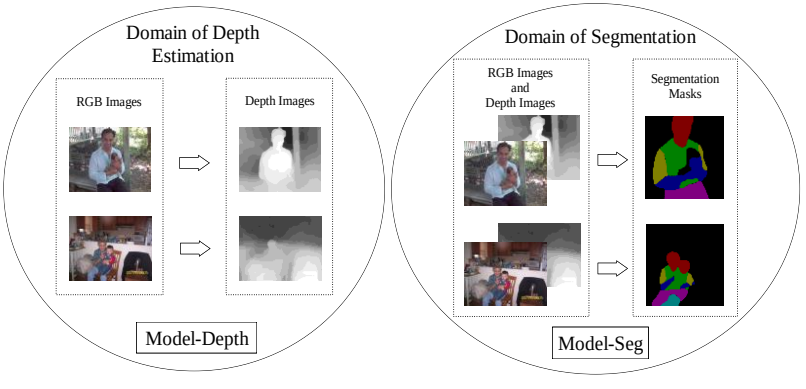


Figure 1: Illustration of the proposed framework. In the left domain, Model-Depth is trained to estimate depth maps from input RGB images. Model-Seg in the right domain learns to conduct semantic segmentation based on both RGB images and the predicted dense depth maps.

The contributions of our work are summarized as follows:

- An approach based on but different from traditional weakly supervised methods is proposed. Through utilizing the knowledge learned from depth estimation, the supervision from automatically obtained depth labels is used to improve the performance of segmentation. This approach solves the problem of limited supervision in semantic segmentation. An improvement of over 5% is achieved without human labeling.

- The correlation between the two tasks, dense depth estimation and semantic segmentation, is explored. Through training based on the initialization provided by Model-Depth, Model-Seg achieves a higher accuracy than when trained only on the dataset for semantic segmentation. The significant improvement on robustness brought by transferring knowledge is shown from Figure 4 to Figure 7.

Related Work

Semantic segmentation and person part segmentation. Semantic segmentation has become an active research topic [7, 9, 18, 19, 20] and the two versions of Deeplab [7, 20] represent the current state-of-the-art. The task-specific features in images have been fully explored and the only way left to improve performance is via data augmentation. Our goal is to avoid the expensive process of data augmentation by transferring the knowledge learned from depth estimation to the segmentation model.

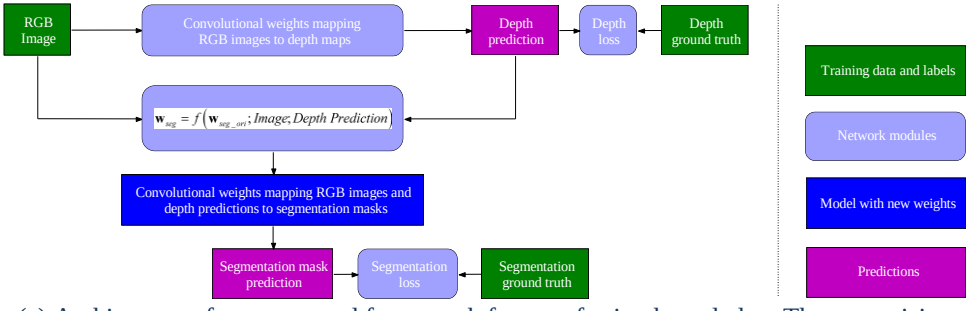
Weakly supervised semantic segmentation. Prior work in this field can be divided into four categories: learning based on bounding boxes, learning based on scribbles, learning based on image tags and mixing multiple types of annotations. Image-level supervision indicates the presence of certain objects in images. Bounding boxes and scribbles indicate the locations and sizes of objects. Point-level supervision indicates the locations and rough boundaries of objects. Two representative works based on bounding boxes are BoxSup proposed in [14] and DeepCut proposed in [21]. The basic idea behind BoxSup is iterating between bounding box generation and the training of convolutional networks. Algorithms based on scribbles include 3D U-Net are proposed in [22] and algorithms based on ScribbleSup are proposed in [16]. 3D U-Net was proposed to perform volumetric segmentation with a semi-automated setup or a fully-automated setup. Algorithms based on image-tags include multi-task learning [23]. Methods based on mixing multiple types of labels utilize image-specific labels together with bounding boxes and scribbles [15, 24]. For datasets with only image-level labels, EM algorithm was used to estimate pixel-level labels. Different from the above-mentioned methods, we introduce another type of weak supervision-depth. The labels are easy to obtain with algorithms such as SfM and MVS.

Task transfer learning. Previous works [25, 26, 27] have tried to predict the parameters in image classifiers from other sources, such as natural language descriptions or few-shot examples. [28] proposed to transform the parameters in detection tasks to segmentation tasks. Few training examples as they requires, an additional model has to be trained to convert the parameters in source domains to those in target domains. Moreover, the training of the additional model heavily relies on the limited available annotations. More related to our work, LSDA proposed in [29] introduced a way to transform image classification parameters into parameters for object detection with limited bounding box annotations. Different from the above methods, the proposed approach transfers knowledge without learning a parameterized function and thus reduces the chance of over-fitting. Model-Depth provides generalizable descriptions of input images and the descriptions are used to augment the feature representations of Model-Seg.

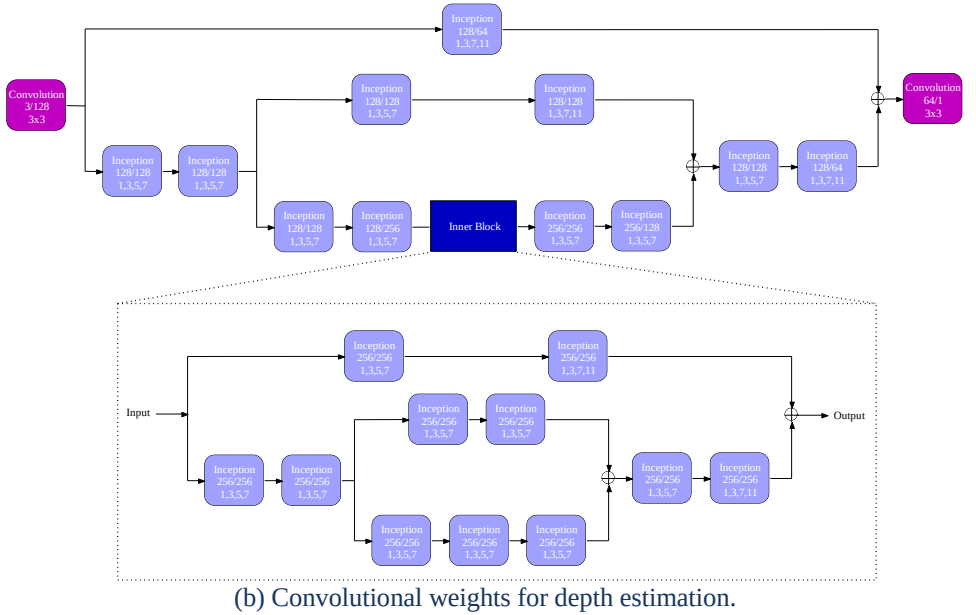
Methodology

In this section we describe the architecture as well as the way of transferring knowledge. The structure is shown in Figure 2 (a). The output tensor from the DCNN for depth estimation, together with input images, are used to transform the original segmentation

model with weights w_{seg_ori} to the new segmentation model with weights w_{seg} . w_{seg_ori} denotes the weights of the original segmentation model which is trained to map RGB images to segmentation masks. w_{seg} denotes the weights of the segmentation model which is trained to map both RGB images and depth predictions to segmentation masks. 3.1 describes the detailed architecture for depth estimation. 3.2 introduces the network for person part segmentation.



(a) Architecture of our proposed framework for transferring knowledge. The supervision from depth labels is used for segmentation.



(b) Convolutional weights for depth estimation.

Figure 2: The proposed framework for semantic image segmentation. (a) The architecture of the framework. (b) The network for estimating depth maps from RGB images.

3.1 Model-Depth for depth estimation

The backbone of the network for depth estimation (Model-Depth) is the Hourglass Network proposed in [30]. The symmetric structure involves convolutions, down-samplings which are followed by up-samplings and convolutions. Figure 2 (b) shows the detailed structure. There are two types of modules: convolutional modules and Inception modules. Each

convolutional module denotes one convolutional layer and each Inception module denotes one Inception unit introduced in [31]. The second line in each convolutional module shows the number of input channels as well as the number of output channels. The third line in each convolutional module shows the sizes of kernels. As is shown in [31], one Inception unit is composed of four convolutional branches, the second line in each Inception module shows the number of input channels as well as the number of output channels. The third line in each Inception module shows the sizes of kernels in the four branches within the module.

The loss function for training the network is the same as the scale-invariant loss introduced in [32]:

$$D(y, y^*) = \frac{1}{n}\sum_{i=1}^{n} d_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} d_i\right)^2 \tag{1}$$

where d_i denotes the difference between the prediction at the $i-th$ pixel and the ground truth label. y denotes the predicted depth map and y^* denotes the ground truth depth map. The large-scale MegaDepth dataset [17] is used to conduct training. The segmentation model with parameters $\mathbf{w}_{seg}$ generalizes better than the original segmentation model with $\mathbf{w}_{seg\_ori}$, as will be shown in Table 1 and 2, Figure 4, 5, 6 and 7.

3.2 Model-Seg for semantic segmentation

The transformation function:

$$\mathbf{w}_{seg} = f\left(\mathbf{w}_{seg\_ori}; Image; Depth\ Prediction\right) \tag{2}$$

converts the original segmentation model with weights $\mathbf{w}_{seg\_ori}$ to another model with weights $\mathbf{w}_{seg}$ based on both the predicted depth maps and input RGB images. Firstly Model-Seg is trained only on the Person Part Segmentation Dataset [9] to develop weights $\mathbf{w}_{seg}$.

The structure of Model-Seg is based on Deeplab-V2 with backbone Resnet-101 [7]. The modification made on Deeplab-V2 is keeping only half of feature channels in the last 12 parallel layers. This is implemented by channel pruning [33]. The reason for pruning the network lies in that it is not necessary to conduct segmentation on such a small dataset with a complex model. The simplified network is able to perform as well as the original complex network.

The method of channel pruning introduced in the paper [33] was based on directly removing channels with smaller L2-norm and re-train the network. However, it is shown in our experiments that after pruning a large portion of channels, it may take a long time for the simplified network to converge. The simplified network even fails to converge after directly removing half of feature channels from critical layers in Deeplab-V2. To address this issue, a scheme is proposed to conduct pruning step by step. In each step, only less than 20% channels are pruned and the network is fine-tuned. The benefits brought by this scheme lie in the guaranteed convergence.

The implementation of (2) is conducted after the training of initial Model-Seg. The initial Model-Seg has to be transformed to conduct segmentation based on both color maps and depth maps. Figure 3 shows Model-Seg as well as the modifications made on the network. As is shown in Figure 3, the new feature representation introduces clues describing the relationship between depth values and semantic predictions. As Model-Depth can be

trained on nearly unlimited data, the predicted depth maps tend to be robust to variances. The knowledge transferred from depth estimations enables Model-Seg to predict semantic masks based on the generalizable depth predictions.

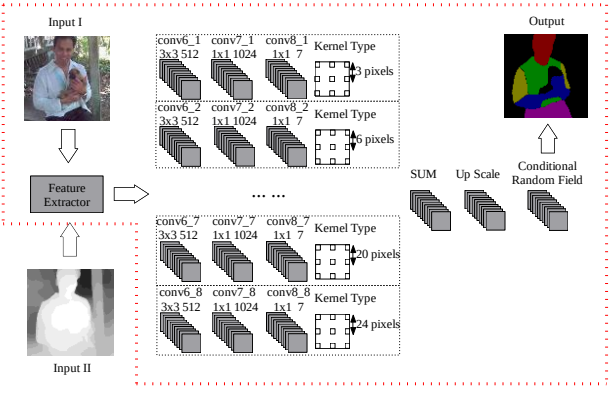


Figure 3: Model-Seg for semantic segmentation. The parts within the red box denote the initial Model-Seg. The part outside the red box shows the prediction from Model-Depth that is added to transform the weights of Model-Seg.

Experiments

4.1 Introduction to the dataset

The dataset for depth estimation is the Mega-Depth Dataset [17]. The dataset involves 196 reconstructed 3D scenes with over 100,000 images. The data for segmentation involves two datasets. The first is the PASCAL VOC 2010 Person Part Dataset [9] for person part segmentation. The Person Part Dataset includes annotations on 3,533 images where 1,716 images are used for training and other 1,817 for test. The second dataset is the LIP (Look into Person) Dataset [34]. The dataset is an order of magnitude larger and more challenging, it contains 50,462 images with elaborated pixel-wise annotations with 19 semantic human part labels. The numbers of images for training, validation and testing are 30,462, 10,000 and 10,000, respectively.

4.2 Implementation details

Firstly, Model-Depth is trained to obtain pixel-level depth predictions. The trained model is used to estimate the depth of images from both the Pascal Person Part Dataset and those from the LIP Dataset. The estimated depth masks are concatenated with original RGB channels to produce the training and testing data for Model-Seg. The backbone of Model-Seg is ResNet-101 with ASPP (Astrous Spatial Pyramid Pooling integrated) [7]. For the Pascal Person Part Dataset, Model-Seg is trained for 40,000 iteration and the batch size is set to 6. For the LIP Dataset, Model-Seg is trained for 400,000 iterations and the batch size is also set to 6. The initial learning rate is set to 2.5e-4 during the training on both datasets.

4.3 Quantitative results

The measure adopted for evaluating segmentation performance is the mean Intersection Over Union (mIOU) proposed in [35]. It is a metric for evaluating semantic segmentation

tasks. It is calculated by dividing the number of true positive samples by the summation of true positive, false negative and false positive samples:

$$mIOU = \frac{1}{N}\sum_{i=1}^{N}\frac{n_{ii}}{t_i + \sum_{j\neq i} n_{ji}} \quad (3)$$

where n_{ji} is the number of pixels of class j which are predicted to class i, and $t_j = \sum_i n_{ji}$ is the total number of pixels belonging to class j. The measure mIOU takes into account both false positives and false negatives. For the Pascal Person Part Dataset, Table 1 shows the comparison of mIOU between the proposed approach and existing solutions. Table 2 shows the performance comparison on the LIP Dataset. The test is conducted on the validation set.

■ Method	mIOU (%)
■ Attention [36]	56.39
■ HAZN [37]	57.54
■ LG-LSTM [38]	57.97
■ Graph LSTM [39]	60.16
■ Deep Lab-V2 (Resnet-101) [7]	64.94
■ Our model shown in Figure 3 (a)	70.19

Table 1: A comparison in mIOU (%) on the Pascal Person Part Dataset [9] between our approach and existing methods.

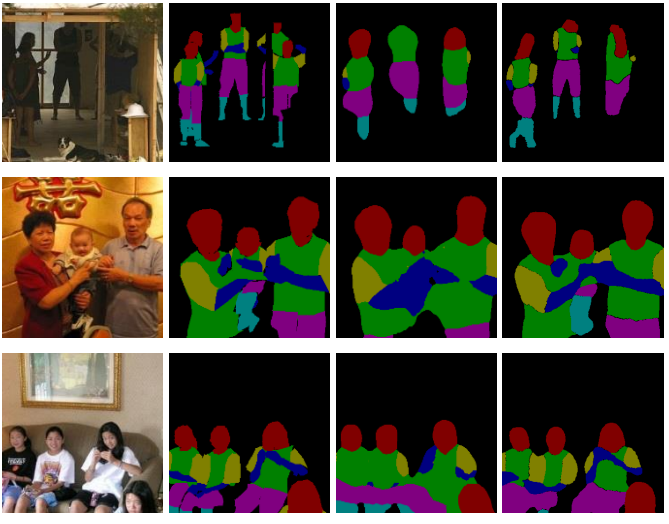
■ Method	mIOU (%)
■ Deeplab-V2 (VGG16) [7]	41.64
■ Attention [36]	42.92
■ Deeplab-V2 (Resnet-101) [7]	44.80
■ Our model	51.08

Table 2: A comparison in mIOU (%) on the Look into Person Dataset [9] between our model and the baseline models.

From Table 1 and Table 2 it can be seen that the proposed method achieves much better performance as well as generalization than the Deeplab-V2 [7] model. The relationship between depth estimation and semantic segmentation contributes to the improvement. The augmented data for depth estimation enables more detailed description of input images with different depth values assigned to different objects. The enriched description facilitates better performance. Actually, the performance shown in Table 2 can still be improved if the optimal training time is found. The improvement highlights the contribution from the knowledge transferred from depth estimation.

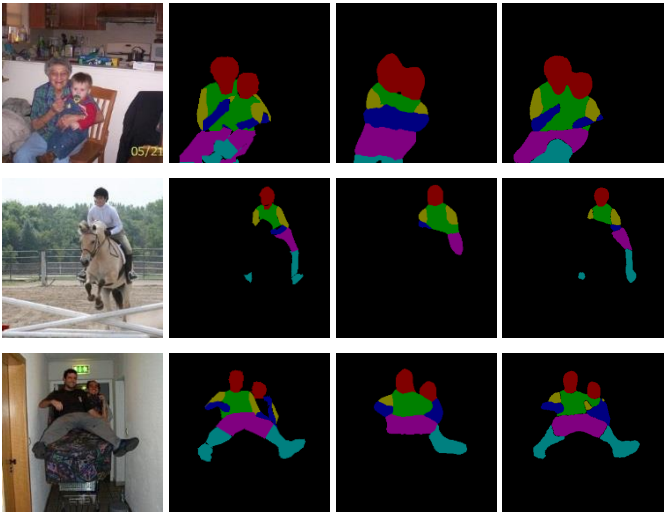
4.4 Qualitative results

Besides quantitative results, some qualitative results are shown below to demonstrate the advantages of the proposed method. Comparison between the performance of the proposed method and that of Deeplab-V2 [7] is conducted on four types of images. In the first type of images, multiple people appear in one image. In the second type of images, the gestures of people are complex. In the third type of images, people only take up small parts in images. In the fourth type of images, people are occluded. Results are shown from Figure 4 to Figure 7. More results will be made available together with our code.



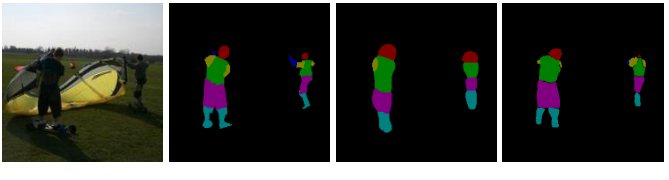
(a) Image (b) Ground Truth (c) Deeplab-V2 (d) Ours

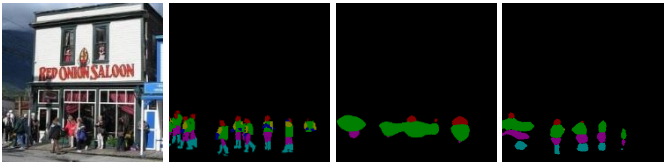
Figure 4: Segmentation results on the PASCAL Person Part Dataset proposed in [9], the input images are with multiple people.



(a) Image (b) Ground Truth (c) Deeplab-V2 (d) Ours

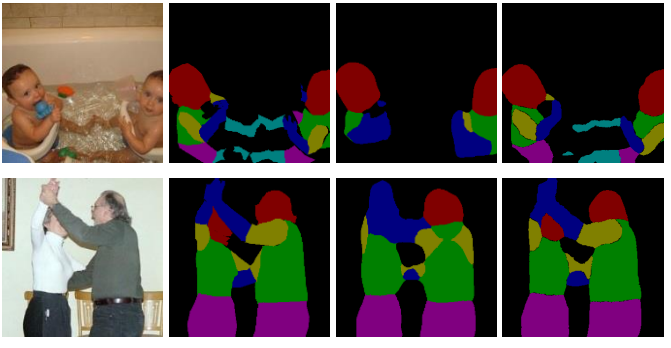
Figure 5: Segmentation results on the PASCAL Person Part Dataset proposed in [9], there are great variances in people's poses.





(a) Image (b) Ground Truth (c) Deeplab-V2 (d) Ours

Figure 6: Segmentation results on the PASCAL Person Part Dataset proposed in [9], people only take up small parts in images.



(a) Image (b) Ground Truth (c) Deeplab-V2 (d) Ours

Figure 7: Segmentation results on the PASCAL Person Part Dataset proposed in [9], people are partially occluded in the images.

Both quantitative and qualitative results have shown that the supervision from depth labels contributes to the improvement on person part segmentation. The improvement is significant.

Conclusion

The proposed approach for person part segmentation integrates the supervision from limited training data with the supervision from depth labels. Model-Seg conducts segmentation based on both color information and depth maps. The depth labels which can be obtained without human labor contribute to the improvement on segmentation. Both quantitative and qualitative results have shown that the proposed scheme significantly improves the performance of person part segmentation on the Pascal Person Part Dataset and that on the LIP Dataset. Furthermore, the strategy of training should be optimized to reduce the training time on the LIP Dataset.

Acknowledgement

This work was partially supported by a research grant from The Hong Kong Polytechnic University (Project Code: 4-BCCJ) and a Natural Science Foundation of China (NSFC) grant (Project Code: 61473243). Mr. Yalong Jiang would like to acknowledge the financial support from PolyU for his PhD study.

References

[1] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740-755.

[2] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., and Zisserman, A., "The pascal visual object classes challenge a retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 198-136, Jan 2015.

[3] M., et al, Cordts, "The cityscapes dataset for semantic urban scene understanding," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213-3223.

[4] Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J., "Path aggregation network for instance segmentation.," in *arXiv preprint arXiv:*, 2018, p. 1803.01534.

[5] Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J., "Large Kernel Matters--Improve Semantic Segmentation by Global Convolutional Network," in *arXiv preprint arXiv:*, 2017, p. 1703.02719.

[6] He, K., Gkioxari, G., Dollár, P., and Girshick, R., "Mask r-cnn," in *arXiv preprint arXiv:*, 2017, p. 1703.06870.

[7] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and L. Alan, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1-1, April 2017.

[8] Bearman A, Russakovsky O, Ferrari V, and Fei-Fei L., "What's the point: Semantic segmentation with point supervision," in *European Conference on Computer Vision*, 2016, pp. 549-565.

[9] Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., and Yuille, A., "Detect what you can: Detecting and representing objects using holistic models and body parts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1971-1978.

[10] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[11] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A., "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, 2017, p. 12.

[12] Snavely, N., Seitz, S.M., and Szeliski, R., "Photo tourism: exploring photo collections in 3D," *ACM transactions on graphics*, vol. 25, no. 3, pp. 835-846, 2006.

[13] Pathak, D., Krahenbuhl, P., and Darrell, T., "Constrained convolutional neural networks for weakly supervised segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1796-1804.

[14] Dai, J., He, K., and Sun, J., "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1635-1643.

[15] Papandreou, G., Chen, L. C., Murphy, K., and Yuille, A. L., "Weakly-and semi-supervised learning of a DCNN for semantic image segmentation," in *arXiv preprint arXiv:*, 2015, p. 1502.02734.

[16] Lin, D., Dai, J., Jia, J., He, K., Sun, J., "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159-3167.

[17] Li, Z., Snavely, N., "MegaDepth: Learning Single-View Depth Prediction from Internet Photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2041-2050.

[18] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical

image segmentation.," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241.

[19] Pohlen, T., Hermans, A., Mathias, M., and Leibe, B., "Full-resolution residual networks for semantic segmentation in street scenes," in *arXiv preprint:*, 2017, p. 1611.08323.

[20] Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation.," in *arXiv preprint arXiv:*, 2017, p. 1706.05587.

[21] Rajchl, M., Lee, M. C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., and Bai, W., "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 674-683, 2017.

[22] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O., "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* , 2016, pp. 424-432.

[23] Vezhnevets, A., and Buhmann, J. M., "Towards weakly supervised semantic segmentation by means of multiple instance nad multi-task learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3249-3256.

[24] Xu, J., Schwing, A. G., and Urtasun, R., "Learning to segment under various forms of weak supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3781-3790.

[25] Ha, D., Dai, A., and Le, Q. V., "Hypernetworks," in *arXiv preprint*, 2016, p. 1609.09106.

[26] Elhoseiny, M., Saleh, B., and Elgammal, A., "Write a classifier: Zero-shot learning using purely textual descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2584-2591.

[27] Wang, Y. X., and Hebert, M., "Learning to learn: Model regression networks for easy small sample learning," in *European Conference on Computer Vision*, 2016, pp. 616-634.

[28] Hu, R., Dollár, P., He, K., Darrell, T., and Girshick, R., "Learning to Segment Every Thing," in *arXiv preprint*, 2017, pp. arXiv:1711.10370.

[29] Hoffman, J., Guadarrama, S., Tzeng, E. S., Hu, R., Donahue, J., Girshick, R., and Saenko, K., "LSDA: Large scale detection through adaptation," in *Advances in Neural Information Processing Systems*, 2014, pp. 3536-3544.

[30] Newell, A., Yang, K., and Deng, J., "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016, pp. 483-499.

[31] Szegedy, C, Liu, W, Jia, Y, Sermanet, P, Reed, S, Anguelov, D, Erhan, D, Vanhoucke, V, and Rabinovich, A, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[32] Eigen, D., Puhrsch, C., and Fergus, R., "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366-2374.

[33] Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J., "Pruning Convolutional Neural Networks for Resource Efficient Inference," in *arXiv*, 2016, p. 1611.06440.

[34] Liang, X., Gong, K., Shen, X., and Lin, L., "Look into Person: Joint Body Parsing & Pose Estimation Network and A New Benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 2018.

[35] Oliveira, G.L., Valada, A., Bollen, C., Burgard, W., and Brox, T., "Deep Learning for human part discovery in images," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1634-1641.

[36] Chen, L.C., Yang, Y., Wang, J., Xu, W., and Yuille, A.L., "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3640 - 3649.

[37] Xia, F., Wang, P., Chen, L.C., and Yuille, A.L., "Zoom better to see clearer: Human part segmentation with auto zoom net," in *arXiv preprint*, 2015, pp. arXiv:1511.06881.

[38] Liang, X., Shen, X., Xiang, D., Feng, J., Lin, L., and Yan, S., "Semantic object parsing with local-global long short-term memory," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3185-3193.

[39] Liang, X., Shen, X., Feng, J., Lin, L., and Yan, S., "Semantic object parsing with graph lstm," in *European Conference on Computer Vision*, 2016, pp. 125-143.