The Second Place Solution for ECCV2022 Vipriors Instance Segmentation Challenge

Fuxing Leng, Jinghua Yan, Peibin Chen, Chenglong Yi

ByteDance,Huazhong University of Science and Technology

Abstract. This paper presents the sencond place solution for eccv2022 vipriors instance segmentation challenge. We build our solution based CBNetV2[4] (using Swin Transformer-Large[6] as backbone). Moreover, we also applied data augmentation policies include Auto Augmentation[7], ImgAug[3], Copy Paste, Horizontal Flip and Multi-scale Training. We demonstrate the accuracy of our method by achieving 50.6 mAP on the test set.

Keywords: Instance Segmentation

1 Method

1.1 Network Architecture

We applied Swin Transformer-Large as backbone, and the pipeline was based on CBNetV2, the pipeline as shown in Figure 1.

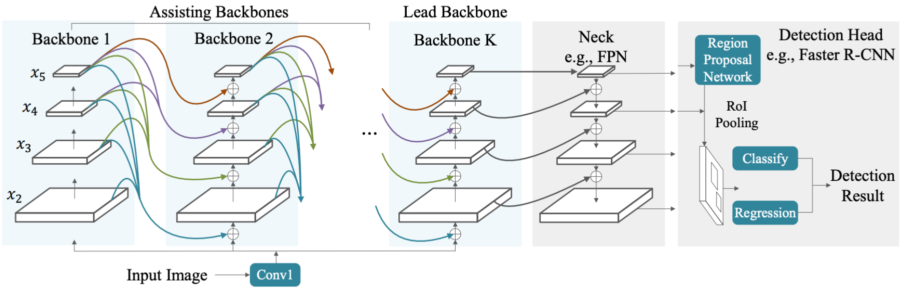

Fig. 1. Our pipeline is the HTC[1] based on the CBNetV2[4]

1.2 Data Augmentation

Data augmentation is a critical component of training deep neural network models. Due to the limited number of training samples, we applied a series of augmentation polices to get rid of over-fitting.

Auto Augmentation: In order to deal with brightness, contrast, color, etc. we applied Auto Augmentation [7] in challenge, which is searched on MS COCO [5]. Notably, we removed all operations on bounding boxes, in order to not change the location of the object mask.

ImgAug: In order to deal with noise and blur, We applied the corresponding data augmentation strategy. Our augmentation policies are implemented by imgaug [3]. Below are the details.

(1). Noise: Gaussian, Shot, Impulse, Speckle.

(2). Blur: Gaussian, Glass, Defocus, Motion.

Copy Paste: We cropped all instance object and mask in advance, then we can paste any instance in the training set during training time, the augmentation as shown in Figure 2.

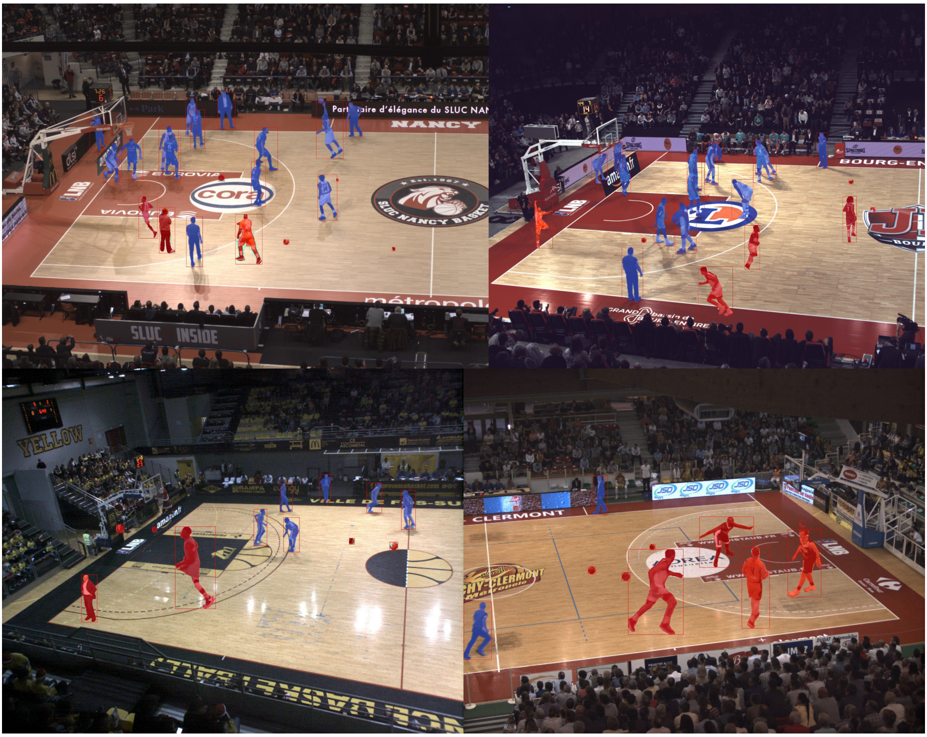

Fig. 2. Online Copy Paste instance object, red instace means copy from other image

2 Experiments

2.1 Experiments Setting

We implement our method using mmdetection [2]. Model was trained using AdamW with a weight decay of 0.05. Following the practice in mmdetection, we adopted multi-scale with horizontal flip augmentation during training. Specifically, we randomly resize the shorter edge of the image within 800 $\sim$ 1400 pixels and keep the longer edge smaller than 1600 pixels without changing the aspect ratio. In inference, we adopted multi-scale testing with image size of 1600×1000, 1600×1400, 1800×1200, 1800×1600, and score threshold of $10e$-3 with Soft NMS.

2.2 Ablation study

As shown in Table 1, Swin-Base achieved 0.221 mAP as baseline with random horizontal flip and multi-scale training. With the data augmentation policies,we achieved 0.362 mAP in the validation set. In the end, we concatenated training set and validation set, with the Swin-Large, we achieved 0.506 mAP in the end. All experiments are given in Table 1.

Table 1. Results of experiments during the challenge

Method	Schedule	AutoAugment	ImgAug	Copy Paste	val set	val mAP	test mAP
Swin-Base	6x					0.221	-
-	+3x	✓				0.286	-
-	+3x	✓	✓			0.294	-
-	+3x	✓	✓	✓		0.362	-
Swin-Large	12x	✓	✓	✓		0.407	-
-	+6x	✓	✓	✓	✓	-	0.506

References

1. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)
2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Jung, A.B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.M., Weng, C.H., Ayala-Acevedo, A., Meudec, R., Laporte, M., et al.: imgaug. https://github.com/aleju/imgaug (2020), online; accessed 01-Feb-2020
4. Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H.: Cbnetv2: A composite backbone network architecture for object detection. arXiv preprint arXiv:2107.00420 (2021)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
6. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
7. Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V.: Learning data augmentation strategies for object detection. In: European Conference on Computer Vision. pp. 566–583. Springer (2020)