INSTANCE SEGMENTATION CHALLENGE TRACK TECHNICAL REPORT, VIPRIORS WORKSHOP AT ICCV 2021: TASK-SPECIFIC COPY-PASTE DATA AUGMENTATION METHOD FOR INSTANCE SEGMENTATION

Jahongir Yunusov
DIVUS corp., Republic of Korea
jahongir7174@gmail.com

Shohruh Rakhmatov
DeltaX corp., Republic of Korea
shoh.mirzo.786@gmail.com

Abdulaziz Namozov
DP World, Republic of Korea
anamozov@gmail.com

Abdulaziz Gaybulayev   **Tae-Hyong Kim**
Department of Computer Engineering
Kumoh National Institute of Technology, Republic of Korea
{g.abdulaziz, taehyong}@kumoh.ac.kr

ABSTRACT

Copy-Paste has proven to be a very effective data augmentation for instance segmentation which can improve the generalization of the model [1]. We used a task-specific Copy-Paste data augmentation method to achieve good performance on the instance segmentation track of the 2nd VIPriors workshop challenge. We also applied additional data augmentation techniques including RandAugment and GridMask. Our segmentation model is the HTC detector on the CBSwin-B with CBFPN with some tweaks. This model was trained at the multi-scale mode by a random sampler on the 6x schedule and tested at the single-scale mode. By combining these techniques, we achieved 0.398 AP@0.50:0.95 with the validation set and 0.433 AP@0.50:0.95 with the test set. Finally, we reached 0.477 AP@0.50:0.95 with the test set by adding the validation set to the training data. Source code is available at https://github.com/jahongir7174/VIP2021.

1 Introduction

Instance segmentation [2] is a useful computer vision task widely used in many real-world applications such as autonomous driving and medical diagnostics. Like other deep learning-based tasks, deep learning-based instance segmentation requires a large training dataset to achieve good performance. Moreover, annotating a large dataset for instance segmentation is more time-consuming than those of other tasks. To address these deep learning issues, the 2nd visual inductive priors for data-efficient deep learning workshop is held at ICCV 2021. Additionally, the workshop offers five computer vision task challenges, including instance segmentation, where models are to be trained from scratch in a data-deficient setting. It is prohibited to use other datasets than the provided training data, so no pre-training and no transfer learning is allowed in every track of the competition.

In the instance segmentation challenge, Synergy Sports, challenge co-organizer, provided a dataset consisting of 184, 62, and 64 images in train, validation, and test sets, respectively. The images are from basketball game scenes, where polygon-based annotations are provided for players and basketball only. In order to achieve satisfactory segmentation accuracy using that small dataset, we tested various state-of-the-art segmentation network models and data augmentation techniques.

The rest of the paper is organized as follows. Section 2 briefly introduces the related work, and section 3 describes the proposed data augmentation method. Then, experiment results are presented in section 4.

2   Related Work

Data Augmentation. General-purpose data augmentation techniques including color-space [3] and geometric [22] transformations are widely used and have shown great classification performance improvements with the ImageNet dataset[5]. Recently, special data transformation methods have been proposed as data augmentation methods, such as Mixup [6] in which two corresponding images are mixed at an arbitrary ratio. Of those methods, Copy-paste [7], which copies object from one image to another, is particularly useful for instance segmentation.

Backbone Network. When constructing deep learning networks for computer vision tasks, backbone models play an important role in determining performance by capturing features from images. Numerous CNN-based backbones were proposed for classification, including AlexNet[8], ResNet[9], ResNeXt[10] and Res2Net[11]. Recently Transformer models, which demonstrated remarkable performances in natural language processing, have been applied to computer vision tasks and have shown succussful results especially in object detection such as Transformer [12], CNN-Transformer DETR [13] and Swin Transformer [14]. CBNetV2 [15], the current latest model, combines high- and low-level features for getting better efficient object detector. This model achieved state-of-the-art performance in both object detection and instance segmentation on the COCO [18] benchmark dataset.

Instance Segmentation. Instance segmentation [15, 9] detects all instances of the target objects and splits the pixels of each detected instance. Hybrid Task Cascade [19] is a recent state-of-the-art model which has a special cascade architecture for instance segmentation.

3   Data Augmentation

Although our data augmentation approach is simple, it is effective in improving the segmentation results and generalizing the instance segmentation model on the given dataset. First, we cropped all objects, then saved the cropped images and their corresponding mask annotations to storage. Then we duplicated images 20 times without any change. All duplicated images are used as input to our copy-paste augmentation method during training. In order to maintain class balance, all class instances are randomly sampled from 5 to 15. Unlike the original copy-paste data augmentation [7], we paste each object, a player or a ball, into a specific location where it should be. The random copy-paste location constraint is as follows:

$$256 \leq x_{min} \leq w - 256 \tag{1}$$

$$\frac{w}{2} + 256 \leq y_{min} \leq \frac{h}{2} + 256 \tag{2}$$

, where w and h are the width and the height of an image, respectively, and (x_{min}, y_{min}) denotes the top left coordinate of an object that is being pasted. Figure 1 shows an example of our data augmentation method.

4   Experiments

In this section, we show the results of some experiments on the provided instance segmentation dataset and compare different methods to verify effectiveness of our method.

The baseline model is the HTC detector [16] on the CBSwin-T backbone with CBFPN [15] using group normalization, trained at the multi-scale mode by a random sampler on the 2x schedule [17]. Note that all ReLU [20] activation functions used in the box and mask heads have been replaced with sigmoid linear unit (SiLU) [21] functions.

In training the original model, the input image is randomly scaled from 400 to 1400 on the short side and up to 1600 on the long side. We adjusted the minimum size of the short side to 800. In our test phase, the single scale (1600, 1400) was used without any test time augmentation (TTA). Except for the last submission, experiments were run on two NVIDIA V100 GPUs. Our last submission used three NVIDIA V100 GPUs.

The detailed step-by-step results of our experiments on validation and test set is described in Table 1. In Table 1, TS Copy-Paste refers to the task-specific copy-paste data augmentation technique described in section 3. In Better backbone setup, we replaced CBSwin-T with CBSwin-B and applied 6x schedule [17]. With the three augmentation methods and the replaced backbone, we achieved 0.398 AP@0.50:0.95 with the validation set and 0.433 AP@0.50:0.95 with the test set. A better test AP might show that the test set has a more similar distribution to the train set than the validation set. Finally, we reached 0.477 AP@0.50:0.95 with the test set by adding the validation set to the training data.
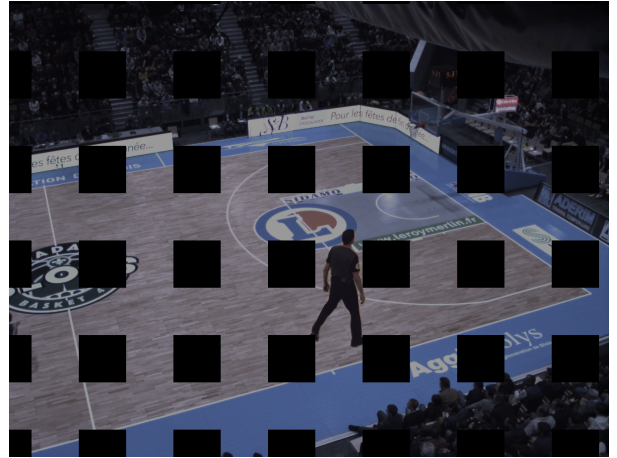
(a) Original input image

(b) Copy-paste data augmentation
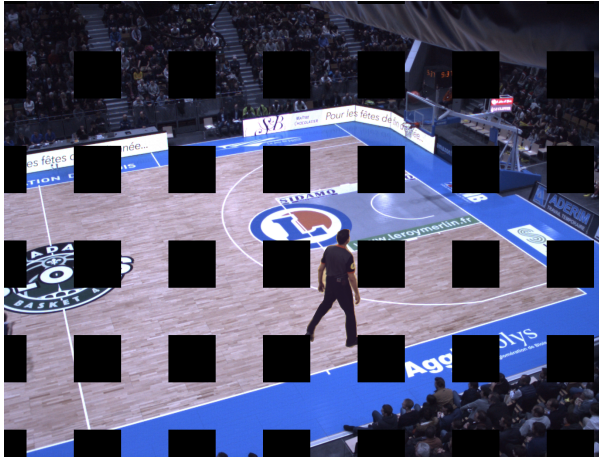
(c) Grid-mask data augmentation

(d) Random data augmentation

Figure 1: Visualization of three data augmentation methods.

Table 1: Step-by-step results on validation and test set

Methods	Schedule	Val AP@0.50:0.95	Test AP@0.50:0.95
Baseline	2x	0.186	-
+ TS Copy-Paste	2x	0.338	-
+ RandAugment [22] + GridMask [23]	2x	0.345	-
+ Better backbone	6x	0.398	0.433
+ Val set added in training	6x	-	0.477

References

[1] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, Barret Zoph. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In CVPR, 2021.

[2] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In CVPR, 2016.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In CVPR, 2015.

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In NeurIPS, 2020

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. In IJCV, 2015

[6] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In ICCV, 2017.

[7] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In ICLR, 2018

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hin-ton. Imagenet classification with deep convolutional neural networks. In NeurIPS, 2012

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[10] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In CVPR, 2017.

[11] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. H. S. Torr. Res2net: A new multi-scale backbone architecture. In TPAMI, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In NeurIPS, 2017.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021.

[15] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi tang, Wei Chu, Jingdong Chen, Haibin Ling. CBNetV2: A Composite Backbone Netwrok Ar-chitecture for Object Detection. arXiv preprint arXiv: 2107.00420, 2021.

[16] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiaping Shi, Wanli Ouyang, Chen Change Loy, Dahua Lin. Hybrid Task Cascade for Instance Segmentation. In CVPR, 2019.

[17] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common ob-jects in context. In ECCV, 2014

[19] Gonlaz Ghiase, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V.Le, Barret Zoph. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In CVPR, 2021

[20] Vinod Nair, Geoffrey E.Hinton. Rectified linear units improve restricted boltzmann machine. In ICML, 2010

[21] Prajit Ramachandran, Barret Zoph, Quoc V. Le, Searching for Activation Function. arXiv preprint arXiv:1710.05951, 2017

[22] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. RandAugment: Practical automated data augmentation with a reduced search space. In ECCV, 2020.

[23] Chen, P., Liu, S., Zhao, H., Jia, J. GridMask data augmentation. arXiv: Computer Vision and Pattern Recognition, 2020.