A Competitive Method for 2021 VIPriors Instance Segmentation Challenge

Bo Yan, Fengliang Qi, Leilei Cao, Hongbin Wang
Ant Group

Abstract

The Visual Inductive Priors(VIPriors) for Data-Efficient Computer Vision challenges ask competitors to train models from scratch in a data-deficient setting. In this paper, we introduce the technical details of our submission to the 2021 VIPriors instance segmentation challenge. Firstly, we designed an effective data augmentation method to improve the problem of data-deficient. Secondly, we conducted some experiments to select a proper model and made some improvements for this task. Thirdly, we proposed an effective training strategy which can improve the performance. Experimental results demonstrate that our approach can achieve a competitive result on the test set. According to the competition rules, we do not use any external image or video data and pre-trained weights. The implementation details above are described in section 2 and section 3. Finally, our approach can achieve 0.402%AP@0.50:0.95 on the test set of 2021 VIPriors instance segmentation challenge.

1. Introduction

Instance segmentation applies widely in image editing, image composition, autonomous driving, etc. Instance segmentation is a fundamental problem in computer vision. Deep learning-based methods have achieved promising results for image instance segmentation over the past few years, such as Mask R-CNN [5], PANet [6], TensorMask [4], CenterMask [11]. SOLO series [9, 10] . The main objective of the VIPriors instance segmentation challenge is to segment basketball players and the ball on images recorded of a basketball court. Different from previous studies, VIPriors instance segmentation challenge does not allow using any pre-trained model, and training data is deficient.

In order to address the problem of data-deficient, we designed an effective data augmentation method, which contains bbox-jitter, grid-mask, hue-transform, color-jitter, random-noise , etc. Then we conducted some experiments to test whether previous studies are effective and made some improvements for this task. Finally, an effective training strategy is proposed to improve the model performance.

2. Approach

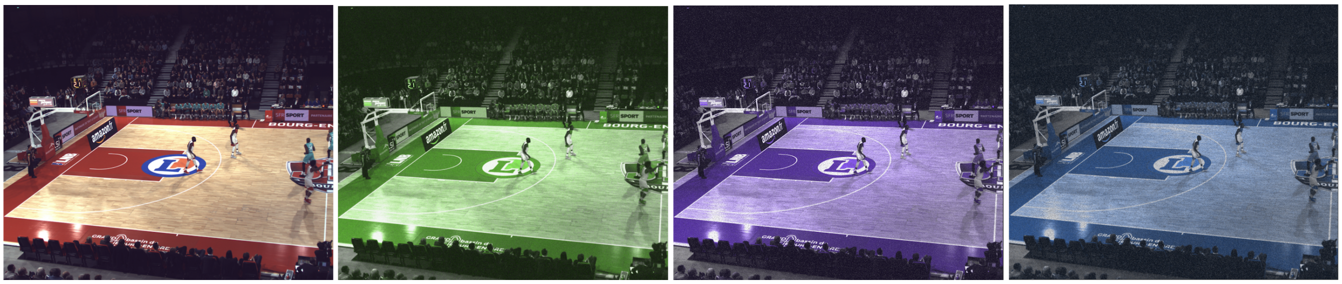
Our approach mainly includes three parts: data augmentation, segmentation model, and training strategy. We introduce the proposed data augmentation strategy in Sec.2.1. The segmentation model is introduced in Sec.2.2. Moreover, the details of training strategy are introduced in Sec.2.3.

2.1. Data Augmentation

For VIPriors instance segmentation, there are 184 images in the training set, 62 images in validation set, and 64 images in test set. Although the scenario seems simple because the task is only to segment basketball players and the ball on images recorded of a basketball court, 184 training images are also insufficient. In order to generate enough images and make the model performs better and more robust, we designed an offline data augmentation strategy that includes four parts, part-1 is color-transform, part-2 is quality-transform, part-3 is filter-transform, and part-4 is hue-transform.

There are four transforms for color-transform: random brightness, random color jitter, random saturation, and random sharpen. We randomly choose one of them as our color-transform to augment the image. There are also four transforms for quality-transform: random blur, random noise, random shuffle pixels, and random pixelization. We also randomly choose one of them as our quality-transform. PIL.ImageFilter is applied to filter-transform. There are five transforms include ImageFilter.DETAIL, ImageFilter.EDGE_ENHANCE, ImageFilter.SMOOTH, ImageFilter.MedianFilter, ImageFilter.ModeFilter, we randomly choose one of them as our filter-transform. Finally, we can change the hue of the image and get images with different colors by hue-transform. Followed by these four transforms pipeline, we can get an augmented image.

We get ten augmented images for every image in training and validation set by the proposed offline augmentation strategy. Finally, our total training images are 2024, and total validation images are 682. Some example augmented



original image augmented image1 augmented image2 augmented image3

Figure 1. Example augmented images by our proposed data augmentation strategy.

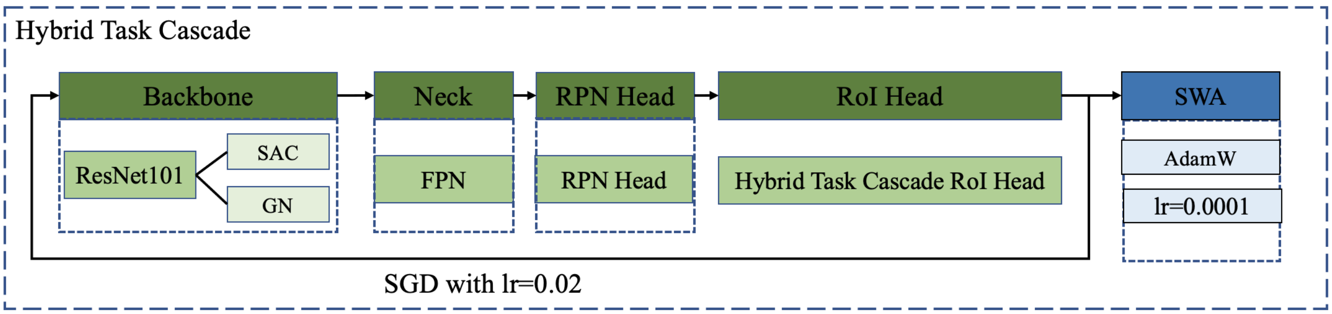


Figure 2. Our model architecture and training pipeline for VIPriors instance segmentation.

images are shown in Figure 1. Some online data augmentation strategies, such as random-flip, random-crop, bbox-jitter, grid-mask [3] are also adopted during training.

2.2. HTC-based Model

Our instance segmentation model is based on Hybrid Task Cascade(HTC) [2]. We use ResNet-101 as our backbone, but we convert the basic convolution layer in the backbone to switchable atrous convolution(SAC) [8], and also we use group normalization [12] to replace batch normalization for better performance.

2.3. Training Strategy

Firstly, to get a proper model for this task, we only use the training set to train the model and validate it on the validation set. Then we train the model with full training and validation set. Stochastic Gradient Descent(SGD) with learning rate(lr)=0.02 is applied to this stage. Finally, SWA [13] training strategy is applied to finetune the model, which can make the model better and more robust. Adam with decoupled weight decay(AdamW) [7] with lr=0.0001 is applied during the SWA training stage.

Our model architecture and training pipeline are shown in Figure 2.

3. Experiments

We train and evaluate our model from scratch without any pre-trained model and backbone, and we do our experiments on a single GPU(NVIDIA Tesla V100 16GB).

3.1. Training Details

We only use the VIPriors instance segmentation training and validation set for training. The total images are 2706 include 2024 training images and 682 validation images after our proposed offline augmentation method. Firstly, we train the model with SGD and lr=0.02. Then the SWA training strategy is applied to finetune the model, and the optimizer is AdamW with lr=0.0001. The images are randomly cropped to [1920, 1440] and then input to the network, and some online augmentation methods are applied include random-flip, bbox-jitter, and grid-mask. The batch size for our training is just 2 because of the limitation of GPU memory.

3.2. Experimental Results

As shown in Table 1, our proposed approach finally achieves 0.402%AP@0.50:0.95 on the VIPriors instance segmentation challenge test set.

3.3. Ablation Study

This section elaborates on how we achieve the final result by ablation study to explain our approach. The baseline is HTC-ResNet101-SAC-GN. We train it only on the training

Methods	Training Data	Epochs	AP@0.50:0.95
HTC-ResNet101-SAC-GN	training-2024	22	35.1%
HTC-ResNet101-SAC-GN + SWA	training-2024	11	37.5%
HTC-ResNet101-SAC-GN	training+validation-2706	20	36.5%
HTC-ResNet101-SAC-GN + SWA	training+validation-2706	10	38.0%
HTC-ResNet101-SAC-GN + SWA + Flip	training+validation-2706	10	38.7%
HTC-ResNet101-SAC-GN + SWA	training+validation-2706	24	38.5%
HTC-ResNet101-SAC-GN + SWA + TTA	training+validation-2706	24	40.2%

Table 1. Comparison results with HTC-ResNet101-SAC-GN on VIPriors instance segmentation test set

set, in order to improve the recall of the model, soft nms [1] is used on the test stage for all experiments. The baseline achieves 35.1% on the test set. Then we use SWA training strategy to finetune this model, after 11 epochs it achieves 37.5% with a significant improvement of 2.4%. Following experiments, we train the model on total images, including training and validation set. HTC-ResNet101-SAC-GN can achieve 36.5%, the score increases to 38.0% after SWA straining 10 epochs, and if we add flip horizontal when test for every image, it can achieve a better score 38.7%. Moreover, we find that if there are more epochs for SWA straining, the score could improve a little, maybe 10 epochs is not enough, and the model has not fitted well. And if we add test time augmentation(TTA) when test, it can brings an improvement of 1.7%. Our test time augmentation includes flip horizontal, multi-scale test with scale factors 0.8, 1.0, and 1.2. Finally, we achieve 40.2% AP@0.50:0.95 on the test set of VIPriors instance segmentation.

4. Conclusion

In this paper, we introduce the technical details of our submission to the 2021 VIPriors instance segmentation challenge, including an effective data augmentation method, a proper model for this task, and an effective training strategy. Experimental results demonstrate that our approach can achieve a competitive result on the test set without external image or video data and pre-trained weights. Finally, our approach achieves 0.402%AP@0.50:0.95 on the test set of VIPriors instance segmentation challenge.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020.

[4] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2061–2069, 2019.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[6] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[8] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10224, 2021.

[9] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.

[10] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *arXiv preprint arXiv:2003.10152*, 2020.

[11] Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, and Lirong Yang. Centermask: single shot instance segmentation with point representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9313–9321, 2020.

[12] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[13] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Swa object detection. *arXiv preprint arXiv:2012.12645*, 2020.