The First Place Solution for CVPR 2023 AVA Challenge - Keypoint Track

Chenglong Yi, Fuxing Leng
Huazhong University of Science and Technology, ByteDance

Abstract

This paper presents the first place solution for CVPR2023 AVA Accessibility Vision and Autonomy Challenge - Keypoint Track. We designed our solution based on Top-down method, which applied CBNetV2 [5] as detector and followed by single-object pose estimator. During the first stage, we applied Swin-Large [6] as CBNetV2 backbone, and some data augmentation policies were also used, including Auto Augmentation [7], Mixup [3], Copy Paste, Horizontal flip and Multi-scale training; during the second stage, we applied VIT-Huge [4] as a strong encoder. As a result, we got 90.96 AP on the test set.

1. Method

Our method overview as shown in Figure 1

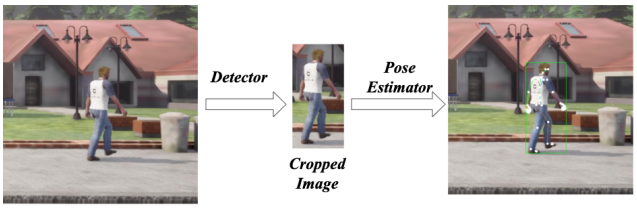

Figure 1. Our method overview

1.1. First Stage

We applied Swin Transformer-Large as backbone, and the pipeline was based on CBNetV2, the pipeline as shown in Figure 2.

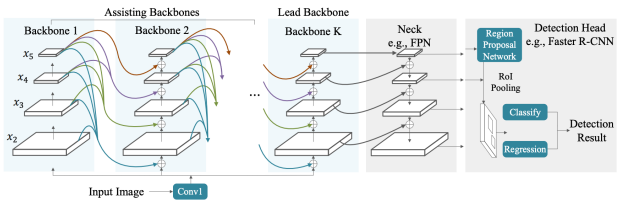

Figure 2. Detector is HTC [2] based on the CBNetV2 [5]

1.2. Second Stage

Top-down methods divide the task into two stages: object detection, followed by single-object pose estimation given object bounding boxes. Instead of estimating keypoint coordinates directly, the pose estimator will produce heatmaps which represent the likelihood of being a keypoint, following the paradigm introduced in Simple Baselines for Human Pose Estimation and Tracking [1]. As shown in Figure 3.

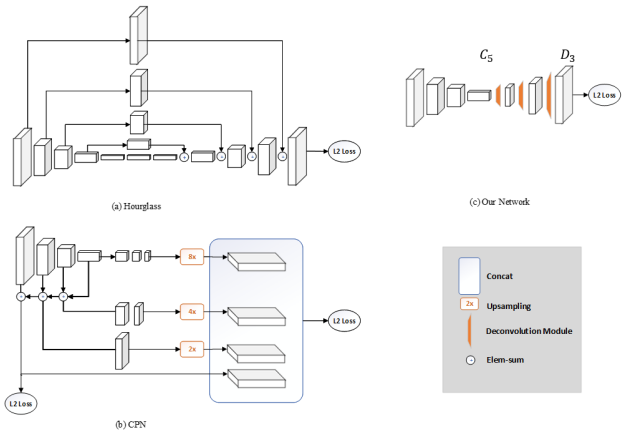

Figure 3. Method of Simple Baselines for Human Pose Estimation and Tracking

2. Experiments

2.1. Experiments Setting

First stage. The detector was trained in 12 epochs. The initial learning rate was $5e^{-5}$, which decayed 0.1 during 8 epochs and 11 epochs. We adopted multi-scale with horizontal flip augmentation during training. Specifically, we randomly resized 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 and score threshold of $1e^{-3}$ with SoftNMS.

Second stage. Got the bounding boxes from the first stage and cropped person image from original image. The

Rank	Method	AP	AP50	AP75
1	**ours**	**90.96**	**95.39**	**92.37**
2	-	83.70	90.82	85.78
3	-	76.64	82.25	78.79
4	-	76.19	82.96	62.73
5	baseline	61.67	81.02	61.97

Table 1. Results of the challenge

cropped image would be resized to 192×256 resolution, and trained pose estimator with 100 epochs. The initial learning rate was $1e^{-4}$, which decayed 0.1 during 60 epochs and 80 epochs.

2.2. Experiments results

As shown in Table 1, based on our top-down method, we achieved competitive results on the test set.

References

[1] Yichen Wei Bin Xiao, Haiping Wu. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision*, pages 466–481. Springer, 2018. 1

[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. 1

[3] Yann N. Dauphin David Lopez-Paz Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2018. 1

[4] Tao Wang Weihao Yu Yujun Shi Zi-Hang Jiang Francis E.H. Tay Jiashi Feng Shuicheng Yan Li Yuan, Yunpeng Chen. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567. IEEE Xplore, 2021. 1

[5] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnetv2: A composite backbone network architecture for object detection. *arXiv preprint arXiv:2107.00420*, 2021. 1

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[7] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *European Conference on Computer Vision*, pages 566–583. Springer, 2020. 1