Self-Constrained Inference Optimization on Structural Groups for Human Pose Estimation

Zhehan Kan, Shuoshuo Chen, Zeng Li, and Zhihai He

Southern University of Science and Technology
{kanzh2021, chenss2021}@mail.sustech.edu.cn
{liz9, hezh}@sustech.edu.cn

1 Complexity Analysis

We chose FCN-VGG [1, 3] as backbone network for our proposed self-constrained inference optimization (SCIO) method. We designed a prediction-verification network to capture the structural relationship between keypoints and optimize the pose estimation result. In the following experiment, we evaluate the impact of the prediction-verification network size on the overall pose estimation accuracy. We trained three prediction-verification networks with different number of model parameters by modifying the size and number of the convolutional layers. The number of parameters and corresponding AP results on the COCO dataset [2] are reported in Table 1. We can see that, when the size of each FCN-VGG backbone is reduced from 14M to 8M, the total number of model parameters is reduced from 168M to 72M and AP decreases by a small margin (1.4%). Compared to the baseline, our method increases the training time by 50% and inference time by 82.4%.

Table 1. Number of parameters and AP for different backbones of refinement network.

	# Params	AP
FCN-VGG	14M	79.2
	8M	78.6
	6M	77.8

2 Study on the Perturbation Term

During the search process of inference, we used perturbations with different step size and chose the best step size of 1.5. Performance results for different perturbation step size are shown in Table 2.

3 Additional Results

In this section, we provide addition results of our proposed SCIO method.

Table 2. AP for different perturbation step size.

Step size	0.5	1.0	1.5	2.0	2.5	3.0	3.5
AP	78.5	78.3	79.2	78.8	78.7	79.0	78.4

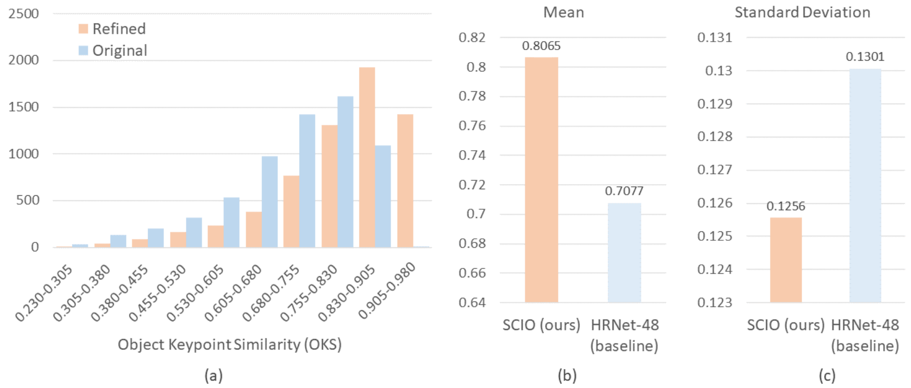

Fig. 1. Comparison between HRNet-W48 (baseline) with SCIO (ours). (a) shows the distribution of pose estimation scores for all persons in COCO val2017. (b) and (c) show mean and standard deviation statistics of pose estimation scores, respectfully.

3.1 Comparison between HRNet-W48 with SCIO

To systematically evaluate our SCIO and study the contribution of each algorithm component better, we use HRNet-W48 [4] as baseline to perform two additional experiments on the COCO val2017 dataset. In Fig. 1, we compare the distribution of OKS scores of all persons in the dataset between our SCIO method with the HRNet-W48 method on COCO val2017 dataset. We can see that our method significantly outperforms the HRNet-W48. Specifically, the number of persons with low pose estimation scores in our method is much smaller than that in the HRNet-W48 method. Meanwhile, the number of persons with high pose estimation scores in our method is much larger. Using our SCIO method, the mean value of the OKS score is increased by about 0.1, and the standard deviation drops by around 0.01.

3.2 Effectiveness of Self-Constrained Optimization (SCO)

Our algorithm has two major components, the self-constrained learning (SCL) and the self-constrained optimization (SCO). Using the verification network as a verification module, the SCO method is able to perform local refinement of the pose estimation result and significantly improve the pose estimation accuracy. In the following experiment, we randomly choose 8 keypoints to demonstrate how the SCO is able to improve their pose estimation accuracy. Specifically, for each keypoint, during the local search, we randomly select 35 keypoints near the

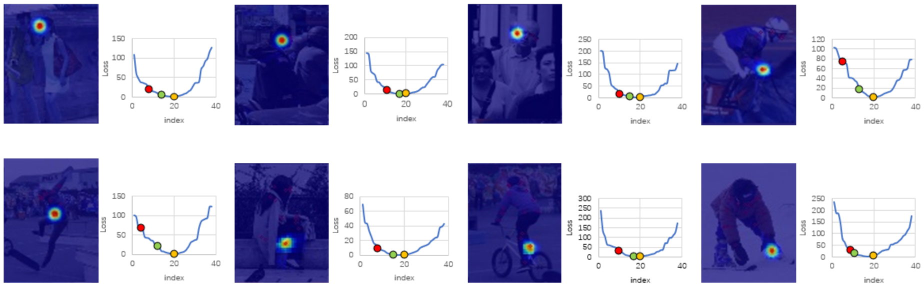

Fig. 2. Eight examples of refined keypoints from SCIO. The distribution shows keypoints error by the baseline HRNet-W48 method (red dots), our method with SCL modules (green dots) and our method with both SCL and SCO modules (yellow dots), where the blue curves represent errors of randomly selected coordinates.

Table 3. Comparison with DARK and Graph-GCNN of input size 128×96 on COCO val2017.

Method	Backbone	Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
DARK [6]	HR48	128×96	71.9	89.1	79.6	69.2	78.0	77.9
Graph-PCNN [5]	HR48	128×96	72.8	89.2	80.1	69.9	79.0	78.6
SCIO (Ours)	HR48	128×96	**73.7**	**89.6**	**80.9**	**70.3**	**79.4**	**79.1**
Performance Gain			**+0.9**	**+0.4**	**+0.8**	**+0.4**	**+0.9**	**+0.8**

original prediction result and compute their self-constrained loss $\mathbf{L_2}$. As shown in Fig. 2, the red dot represents keypoint loss (or estimation error) by the baseline HRNet-W48 method, the green dot represents keypoint error by our method with SCL modules, and the yellow dot represents final keypoint error with both SCL and SCO modules. We can see that the keypoint error decreases gradually when our method is used and both modules are contributing significantly to the overall performance.

3.3 Comparison on Different Input Size

Table 3 shows the performance comparison on pose estimation with different input image size, for example 128×96 instead of 384×288. We have only found two methods that reported results on small input images. We can see that our SCIO method also outperforms these two methods on small input images.

References

1. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR. pp. 7103–7112 (2018)

2. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
4. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
5. Wang, J., Long, X., Gao, Y., Ding, E., Wen, S.: Graph-pcnn: Two stage human pose estimation with graph pose refinement. In: ECCV. pp. 492–508 (2020)
6. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: CVPR. pp. 7091–7100 (2020)