Progressive Target Refinement by Self-Distillation for Human Pose Estimation

Jingtian Li, Lin Fang, Yi Wu, and Shangfei Wang

University of Science and Technology of China
sfwang@ustc.edu.cn

Abstract. The handcrafted heatmap target can be improved and one way is knowledge distillation, which takes the predicted heatmaps from another model as auxiliary supervision. However, previous pose distillation methods are training inefficient, requiring either an extra training stage or complex network architecture modification. In this paper, we propose a novel **S**elf-**D**istillation for Human **P**ose Estimation (SDP) method for better distillation efficiency. Specifically, a student pose estimator distills the soft targets from itself with the backup information of a previous batch, where the targets are progressively refined through model updating. The main advantage of our method is that we achieve efficient training and simple implementation simultaneously. Existing pose estimation networks can benefit from the proposed method effortlessly. A stepping strategy, that widens the distillation distance with the decaying of the learning rate, is further proposed. It ensures the difference between teacher and student in a low learning rate condition. Experimental results on two widely-used benchmark datasets, MPII and COCO, illustrate the effectiveness of the proposed approach.

Keywords: heatmap · pose Estimation · knowledge distillation

1 Introduction

Human Pose Estimation aims at predicting where the human keypoints locate in a given image. It is a fundamental computer vision technique and is widely applied in downstream tasks, such as action recognition and virtual reality, etc. Based on the different representations of targets, previous methods can be divided into two categories, i.e., coordinate regression and heatmap regression. The heatmap representation achieves better performance than the coordinate representation because of rich spatial supervision. Therefore, a large number of previous works were based on the heatmap representation [2, 4, 14, 17, 19].

In current practice, the Gaussian distributions of heatmap targets are assigned the same scales and confidences for all samples. This is unreasonable since the uncertainty varies for different input images and keypoints. For instance, the occluded keypoint samples hold higher uncertainty than visible ones. The keypoints of a person with rare postures, like squatting, also need larger activation areas and lower confidence than other keypoints.

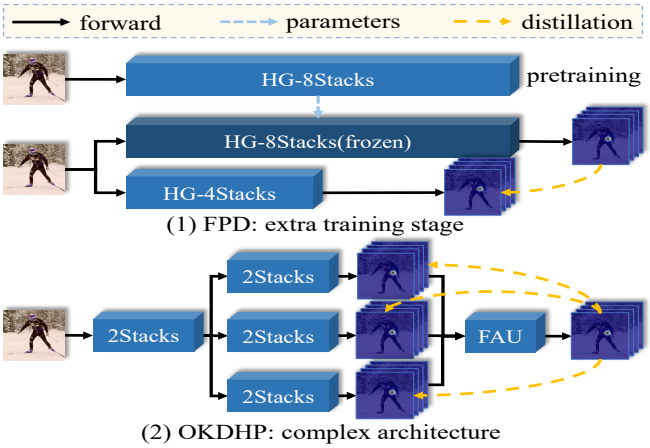

Fig. 1: Illustration of the limitations of previous pose distillation methods.(1) FPD transfers the knowledge from a pretrained teacher model, requiring an extra training stage. (2) OKDHP distills the multi-branches from each other, where the original network is modified into a complex architecture.

A promising approach to get adaptive heatmap targets is knowledge distillation. The predictions from the teacher model provide refined supervision for the student model [3, 8]. Recently, some works applied pose distillation to improve the performance of lightweight models, as presented in Figure 1. Conventional offline distillation was first applied in Fast Pose Distillation (FPD) [21], which takes the heatmap predictions from a large teacher model as soft targets and distills a light student pose estimator. The pose distillation process includes a stage of pretraining the teacher model and a knowledge transfer stage. Such a two-stage manner is parallelization inefficient. Online Knowledge Distillation framework by distilling Human Pose(OKDHP) [10] simplified the distillation procedure to one stage by introducing online knowledge distillation. In specific, multiple peer network branches are developed to provide soft supervision for each other. However, such an online distillation design requires complex network architecture modifications, resulting in hard implementation.

To improve the heatmap targets with better efficiency, we propose a novel **S**elf-**D**istillation for Human **P**ose Estimation (SDP) method, in which the pose estimator distills from itself. Specifically, we transfer the prediction generated in a previous batch as the soft target, which is progressively refined with the model updating. Furthermore, to ensure the difference between the teacher and the student, we propose a stepping strategy that widens the distillation distance when the learning rate degrades. With a smaller learning rate, we transfer the soft target from a batch with more iterations away instead of the batch neighboring. Our method improves both training stability and model performance as a consequence of the progressive refinement of heatmap targets. Moreover, compared with existing pose distillation approaches, SDP has only one training stage

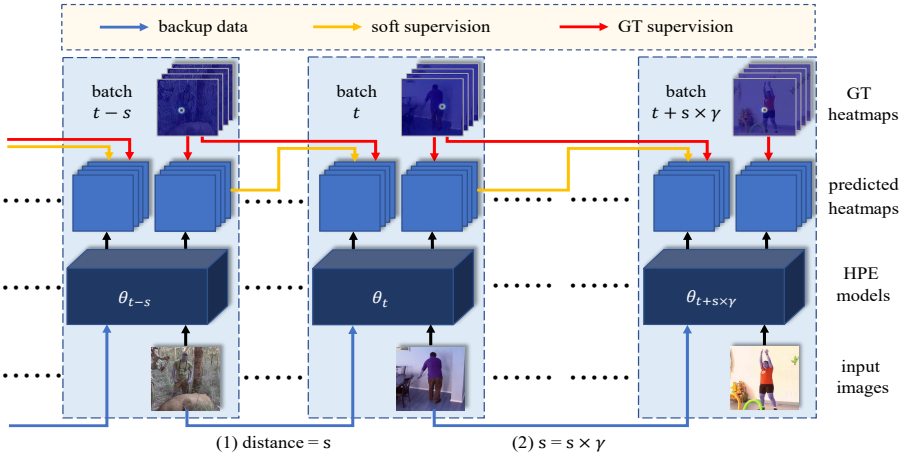

Fig. 2: Overview of the proposed Self-Distillation for Human Pose Estimation method.

and is free from network architecture modification. Therefore, it is effortless for existing pose estimation models to benefit from the proposed method.

The main contributions of this paper are as follows.

- To our best knowledge, we are the first to propose the self-distillation pose estimation method. It is training efficient and implementation simple simultaneously, compared with previous pose distillation methods.
- We further design a stepping strategy to ensure the difference between the teacher and the student in a low learning rate condition, boosting self-distillation efficiency.
- Comprehensive experimental results on two widely-used benchmark datasets, MPII and COCO, validate the effectiveness of the proposed method.

2 Related Work

In the past years, heatmap-based pose estimation gained great progress after first being introduced in [18]. Thereafter, several network architectures were proposed thereafter in seizing of high performance. Stacked Hourglass [12] architecture consists of several repeated hourglass-shaped modules, each of which starts with a downsampling process and then upsampling. Simple and effective baseline methods called Simple Baselines [19] are proposed based on ResNet [7]. HRNet [15] maintains high-resolution representations through the whole feature extraction process to maximize the preservation of precise location information.

Although heatmap-based pose estimators have achieved remarkable performance through network architecture iterations, the current practice of heatmap representation still needs improvement. It's unreasonable to generate the ground-truth Gaussian distribution with the same scales and confidences for all samples.

Pose distillation is a promising approach to refine the heatmap targets. It was exploited in some previous methods for lightweight pose estimation. FPD [21] applied the conventional offline distillation, which requires two stages for training and thus is parallelization inefficient. To simplify the distillation procedure, OKDHP [10] developed multiple branches and performed distillation by learning from each other in a one-stage online manner. However, such an online distillation design is hard to implement due to complex network architecture modifications. To cope with the limitations of these methods, we propose a novel self-distillation pose estimation method. The knowledge for the student is transferred from an old version of itself dent at the same time. Peer students and a network with multiple branches were developed to provide knowledge for each other. Although online distillation simplifies the procedure of knowledge transfer, it is realized with the cost of heavy network architecture modification.

Furthermore, unlike those knowledge distillation methods with discriminative knowledge, we instead transfer the dense heatmap targets with spatial pixel-wise information, which is more challenging. Moreover, to our best knowledge, previous self-distillation methods transfer the knowledge with a fixed distillation distance, such as an epoch, a batch, etc. Such approaches are unsatisfactory because distillation efficiency degrades with a decaying learning rate. Our method proposes to widen the distillation distance when the learning rate degrades, which ensures the difference between the teacher and the student.

3 Method

In this section, we present the proposed Self-Distillation for Human Pose Estimation (SDP) Method, which is composed of the self-distillation framework and the stepping strategy, as shown in Figure 2.

3.1 Self-Distillation

We propose gradually improving the heatmap objectives in a self-distillation manner to get over these restrictions. Instead of transferring the knowledge from another complex teacher model, we utilize the backup information from a previous batch for pose distillation. As shown in Figure. 2, each batch of training has two columns. Taking the t-th batch for example, the left column stands for the backup images $\mathbf{I}^{t-s}$ transferred from the (t-s)-th batch, while the right column denotes the randomly sampled images $\mathbf{I}^{t}$. Both of them are fed into the model:

$$\begin{aligned} \hat{\mathbf{H}}_t^t &= \mathbf{\Theta_t}(\mathbf{I}^t) \\ \hat{\mathbf{H}}_t^{t-s} &= \mathbf{\Theta_t}(\mathbf{I}^{t-s}) \end{aligned} \tag{1}$$

where $\mathbf{\Theta_t}$ is the model of the t-th iter, $\hat{\mathbf{H}}_t^t$ and $\hat{\mathbf{H}}_t^{t-s}$ are the heatmaps predicted by the model $\mathbf{\Theta_t}$ to the images $\mathbf{I}^t$ and $\mathbf{I}^{t-s}$.

These predictions are first supervised by the ground truth heatmap target, which is the top row in Figure 2:

$$\mathcal{L}_{gt} = MSE([\hat{\mathbf{H}}_t^t, \hat{\mathbf{H}}_t^{t-s}], [\mathbf{H}^t, \mathbf{H}^{t-s}]) \tag{2}$$

where $\mathbf{H}^t$ and $\mathbf{H}^{t-s}$ are the ground truth heatmap target for images $\mathbf{I}^t$ and $\mathbf{I}^{t-s}$, and $[,]$ denotes the concatenate operation. $\mathcal{L}_{gt}$ stands for the supervision from the ground truth target, pictured as the gray line in the figure, which needs improvement.

The soft predictions from the (t-s)-th batch are also backed up as the refined target, with which the self-distillation loss is measured as follows:

$$\mathcal{L}_{sd} = MSE(\hat{\mathbf{H}}_t^{t-s}, \hat{\mathbf{H}}_{t-s}^{t-s}) \tag{3}$$

where $\hat{\mathbf{H}}_{t-s}^{t-s}$ is the heatmap predicted by the model $\mathbf{\Theta_{t-s}}$ to the images $\mathbf{I}^{t-s}$, and $\mathcal{L}_{sd}$ stand for the supervision from the distilled soft target, pictured as the red line in the figure. With the model being updated through training data, the soft heatmap target $\hat{\mathbf{H}}_{t-s}^{t-s}$ is progressively refined, leading to better performance.

3.2 Stepping Strategy

Previous pose estimation methods usually reduce the learning rate by a multiplicative factor after each pre-defined milestone [15]. Such a learning rate scheduler is supposed to stable the training process and help the model converge to a better performance. However, when the learning rate decays, the heatmap prediction $\hat{\mathbf{H}}_t^{t-s}$ and the soft target $\hat{\mathbf{H}}_{t-s}^{t-s}$ in Eq.(3) have less difference with a fixed s. In such circumstances, the transferred knowledge is poor, and the distillation efficiency degrades [3, 8].

To overcome this problem, we propose a stepping strategy to ensure the difference between the teacher and the student. Specifically, we denote the hyper-parameter s as the *distillation distance*, which is widened as the learning rate decays. As shown in Figure 2.(2), when the training comes to a preset milestone, the learning rate lr and the distillation step s are updated simultaneously as follows:

$$\begin{aligned} lr &= lr \ / \ \gamma \\ s &= s \times \gamma \end{aligned} \tag{4}$$

where γ is a multiplicative factor, usually set to 10 in practice. With the proposed stepping strategy, the student is able to pick a teacher with more different knowledge, regardless of the slow model updating.

3.3 Overall Loss

In summary, taking Eq.(2) and Eq.(3) together, the total objective function of the proposed method consists of a conventional loss L_{gt} from the ground truth target and a self-distillation loss L_{sd} from the distilled soft target:

$$\mathcal{L}_{total} = \mathcal{L}_{gt} + \alpha \times \mathcal{L}_{sd} \tag{5}$$

where α is a hyper-parameter to weight the self-distillation loss.

Table 1: Comparison with the state-of-the-art methods on the MPII test set. Metric: PCKh@0.5.

Method	Head	Sho.	Elb.	Wri.	Hip	Kne.	Ank.	Mean
Newell *et al.* [ECCV'16] [12]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ning *et al.* [TMM'17] [13]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu *et al.* [CVPR'17] [6]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chen *et al.* [ICCV'17] [5]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.* [ICCV'17] [20]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al.* [ECCV'18] [9]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang *et al.* [ECCV'18] [16]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
FPD [CVPR'19] [21]	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
OKDHP [CVPR'21] [10]	98.2	96.6	92.3	88.0	91.0	88.5	84.5	91.7
SDP(Ours)	98.4	96.9	92.7	88.6	91.3	89.0	85.5	**92.1**

4 Experiments

4.1 Experimental Conditions

DataSet We conducted experiments on two popular datasets, MPII [1] and COCO [11]. The MPII dataset has around 25K images with 40K subjects labeled with 16 keypoints. We followed the train / valid / test split strategy in [18]. The COCO dataset involves over 200,000 images and 250,000 person instances labeled with 17 keypoints. We followed the commonly used train2017 / val2017 / test-dev2017 split for training and evaluation.

Training We follow the training details in previous pose distillation methods [10, 21], including input size, data augmentation, total epochs, and learning rate schedule. For MPII, the input is in 256×256 pixels size, while we resize the cropped image to 256×192 pixels for COCO. The augmentation includes horizontal flipping, random rotation, random scaling, and half-body data augmentation. The loss weighting factor α in Eq.(5) and the initial distillation distance s in Eq.(4) are set to 3 and 1. Experiments were done with Pytorch1.10 on a Ubuntu18.04 server, which has 4 GeForce RTX 3090 GPUs.

Testing To get the bounding boxes of each person when testing, we utilized the official detection results for the MPII dataset. For the COCO dataset, we use the same person detectors provided by Simple Baseline [19] for both the val2017 set and the test-dev2017 set. We applied a flip test for both datasets and a six-scale pyramid testing procedure for the MPII test set.

Evaluation Metric We evaluate the model performances based on the OKS (object keypoint similarity) score for the COCO dataset and the PCKh (head-normalized probability of correct keypoint) score for the MPII dataset, following previous works [10, 21]. For the COCO dataset, we report standard average precision (AP) and average recall (AR) scores. For the MPII dataset, we use the PCKh@0.5 ($\alpha = 0.5$) score. The PCKh score means a joint is correct if it falls within $\alpha \times l$ pixels around the ground truth position, where α is a constant and l is the head size that corresponds to 60% of the diagonal length of the ground-truth head bounding box.

Table 2: Evaluation of the proposed SDP method on COCO val2017 set, with a detector having human AP of 56.4 on COCO val2017 dataset. The major metric is the third column denoted as AP.

SDP	Network	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
✗	SBL-R50	70.4	88.6	78.3	67.1	77.2	76.3	92.9	83.4	72.1	82.4
✓		**71.7**	**89.2**	**79.0**	**68.2**	**78.6**	**77.4**	**93.3**	**84.0**	**73.1**	**83.6**
✗	HG-4Stacks	73.0	89.1	79.8	69.6	79.8	78.5	93.2	84.5	74.3	84.6
✓		**74.2**	**90.3**	**80.5**	**70.6**	**80.8**	**79.4**	**93.7**	**85.9**	**75.3**	**85.7**
✗	HRNet-W32	74.4	90.5	81.9	70.8	81.0	79.8	94.2	86.5	75.7	85.8
✓		**75.2**	**90.6**	**82.3**	**71.5**	**82.2**	**80.5**	**94.2**	**86.8**	**76.3**	**86.7**

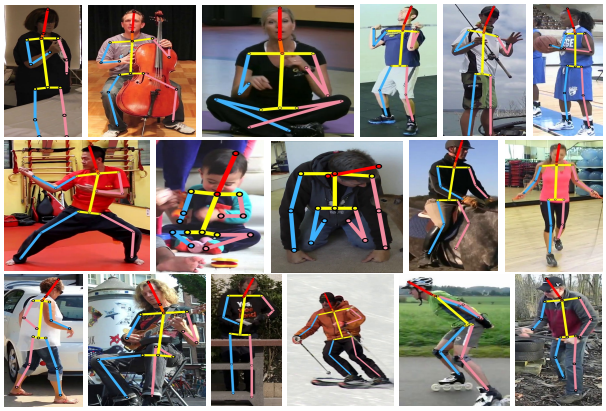

Fig. 3: Visualized results on the MPII dataset.

4.2 Results on MPII dataset and COCO dataset

We evaluated the proposed SDP method on the MPII dataset. Table 5 presents the results of three different-architecture networks, i.e., HG-4Stacks, SBL-R50, and HRNet-W32, trained without and with our method, which are all tested on the MPII valid set. SBL-R50 and HG-4Stacks stand for Simple Baseline with ResNet50 backbone and Hourglass with 4 stacks. The result of the SBL-R50 and HRNet-W32 without SDP are from their original papers, while that of the HG-4Stacks is from the model we trained ourselves since there are no pretrained models from previous works. Table 1 compares the PCKh@0.5 accuracy results of state-of-the-art methods with our method on the MPII testing set. We also provide visualized results in Figure 3 . Besides, we evaluated the effect of the proposed method on the MS COCO keypoint dataset with three popular pose estimation networks. We compare the results of the networks without and with our method in Table 2. The result of the HG-4Stacks without SDP is from the model we trained ourselves since no pretrained models are provided in previous works, while those of the other two networks are from their original papers.

We can observe from the results in Table 5 and Table 2 that, compared to independent training with ground truth labels, the proposed SDP method brings better performances to the networks regardless of their architecture. This

Table 3: Comparison of the distillation efficiency with previous pose distillation methods. The student network is HG-4Stacks for FPD and Ours, while the teacher network for FPD is a larger HG-8Stacks. The training time is rounded.

Method	PCKh@0.5	GFLOPs	Training Time
HG-4Stacks	89.2	14.3	12h
FPD [21]	89.7	66	51h
OKDHP [10]	90.0	47	/
Ours	89.9	**14.3**	**23h**

is because the proposed method provides refined soft supervision, which largely alleviates the limitations of the ground truth label. The benefit from our method on the HRNet-W32 network is smaller than the other two networks mainly because the performance is already high and almost saturated, making further enhancement hard. Especially given that previous state-of-the-art pose estimators improved the PCKh scores by only 0.1% to 0.4% in Table 1, the performance improvements we achieved are remarkable.

Table 4: Ablation study of the stepping strategy on the MPII valid set with an SBL-R50 network. Metrics: PCKh@0.5 and PCKh@0.1.

Stepping	PCKh@0.5	PCKh@0.1
Baseline	88.53	33.91
✗	89.12	34.28
✓	**89.43**	**35.05**

The results in Table 1 illustrate that our SDP method achieves state-of-the-art performance compared to previous methods. Note that the network of our method is the same HG-4Stacks as FPD and OKDHP. We still achieve comparable performance when compared to other top performers with the latest architecture. Overall, the experiments on the MPII dataset verified the effectiveness of the proposed method.

We compare the distillation efficiency of our method with previous pose distillation approaches, i.e., FPD [21] and OKDHP [10]. Metrics including PCKh@0.5, GFLOPs, and Training Time are reported in Table 3. The GFLOPs metric stands for Giga Floating Point of Operations, measuring the computation cost of the network. The results for previous methods are from the paper of OKDHP. We estimated the training time for FPD and our method on the same GPU with no other task running. The training time of OKDHP is not reported because the officially released code is not runnable. Experiments were done on the MPII dataset.

What stands out in the table is the much lower computation cost of our method than those of previous methods. The main reason is that FPD needs an extra teacher model and OKDHP requires peer branches to provide soft targets,

Table 5: Evaluation of the proposed SDP method on the MPII valid set. Metric: PCKh@0.5.

SDP	Network	Head	Sho.	Elb.	Wri.	Hip	Kne.	Ank.	Mean
✗	SBL-R50	96.4	95.3	89.0	83.2	88.4	83.9	79.6	88.5
✓		**97.0**	**95.5**	**89.9**	**85.1**	**88.8**	**85.0**	**81.2**	**89.4**
✗	HG-4Stacks	96.6	95.6	89.1	83.8	88.7	84.9	81.0	89.1
✓		**96.9**	**95.7**	**90.5**	**85.6**	**89.6**	**86.0**	**81.7**	**89.9**
✗	HRNet-W32	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
✓		**97.2**	**96.2**	**91.0**	**86.7**	**89.6**	**87.2**	**83.4**	**90.6**

bringing additional computation. In contrast, the pose estimator of our method learns from itself without backup data, and thus has no additional computation apart from the original network.

Another observation from the table is that our method is faster to train than FPD. It is majorly because FPD needs to train a larger teacher beforehand, and then transfer the knowledge from the teacher to the student. The training of the heavyweight teacher and the distillation stage both take a lot of time. Our method is implemented in a one-stage self-distillation manner, where the distillation computation has high parallelization. Overall, these results indicate that, compared with FPD and OKDHP, the proposed method achieves better training efficiency while keeping better or comparable performance.

4.3 Ablation Study of Stepping Strategy

We conducted an ablation study on the stepping strategy presented in Eq.(4), which widens the distillation step s when the learning rate decays. The experiments were done on the MPII valid set with a Simple Baseline based on the ResNet50 backbone (SBL-R50) network. The loss weighting factor α is set to 3 in this ablation study. The PCKH@0.5 and PCKh@0.1 results are reported in Table 4. As shown in the table, we can see that, although the self-distillation framework already improves the performance greatly, the stepping strategy further boosts the network performance by 0.31% and 0.77% in PCKh@0.5 and PCKh@0.1. This illustrates the importance of the proposed stepping strategy, which ensures the difference between the teacher and the student in pose distillation.

Table 6: Ablation study of the initial distillation distance s on MPII valid set with an SBL-R50 network. Metrics: PCKh@0.5 and PCKh@0.1.

Initial s	1	10	100
PCKh@0.5	**89.43**	89.24	88.77
PCKh@0.1	**35.05**	34.52	33.98

Table 7: Ablation study of the loss weighting factor α on MPII valid set with an SBL-R50 network. Metrics: PCKh@0.5 and PCKh@0.1.

α	0	1	2	3	4
PCKh@0.5	88.53	89.38	89.40	**89.43**	89.37
PCKh@0.1	33.91	34.96	34.85	**35.05**	35.00

4.4 Influence of different initial distillation distance

Furthermore, We experimented with different settings of the initial distillation distance s on the MPII valid set with an SBL-R50 network. Table 6 presents the performance results, which indicates that the performance degrades with an order of magnitude increase in the distillation distance. Such results indicate that it is suitable to set a large initial s. The probable reason is distilling with a long distillation distance slows down the convergence in the beginning epochs of training.

4.5 Ablation Study of Loss Weight

We conducted an ablation study on the loss weighting factor α in Eq.(5), which balances the supervision from the ground truth target and the soft target. The experiments were done on the MPII valid set with a Simple Baseline based on the ResNet50 backbone (SBL-R50) network. The initial distillation s is set to 1 in this ablation study. Table 7 provides the experimental results.

As shown in the table, we have three major observations. First, there is a performance gap between the networks with α set to 0 and 1, from 88.53 and 33.91 to 89.38 and 34.96 in PCKh@0.5 and PCKh@0.1. This indicates the remarkable improvement from the self-distillation loss term. The main reason is our method provides soft targets by self-distillation, which is further progressively refined with the model updating. Second, the network performance roughly follows the trend of first increasing and then decreasing. It achieves a peak performance of 89.43 PCKh@0.5 and 35.05 PCKh@0.1 when *alpha* is set to the optimal value of 3. These results suggest that the distillation loss term is probably more important than the ground truth loss. Third, on the other hand, we discovered that this hyperparameter setting is not sensitive and has a wide range of acceptable values. The performance is robust regardless of α varying from 1 to 4, which also indicates the effectiveness of the proposed method.

4.6 Target Visualization

Figure 4 visualizes the refined heatmap target on the MPII dataset. The heatmaps in the figure are a α-weighted average of the ground truth heatmaps and the soft heatmaps. We have two major observations from the figure. First, lower confidences and larger activation areas are assigned to the heatmap targets for the

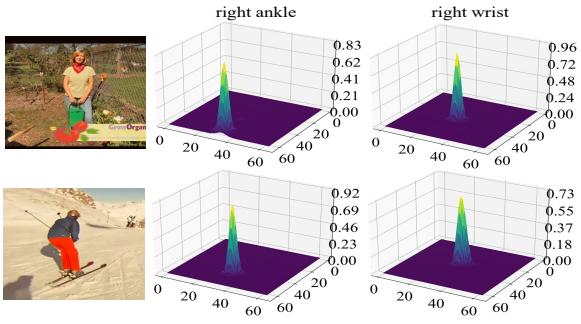

Fig. 4: Visualization of the soft heatmap targets on the MPII dataset.

occluded keypoints than visible ones. In the first row, the max of the heatmap target for the right ankle is 0.13 smaller than that for the right wrist. The same goes for the image and heatmaps in the second row, where the heatmap target of the right wrist holds a maximum of 0.19 lower than that of the right ankle. Second, compared to the first image with a normal standing gesture, the target confidences are lower for the second image. For occluded keypoints, the comparison between the two images is 0.73 against 0.83, while for visible keypoints, it is 0.92 versus 0.96. In summary, the figure shows that our SDP succeeded in assigning heatmap targets with adaptive activation area and confidence.

5 Conclusion

In this paper, we propose a novel Self-Distillation for Human Pose Estimation (SDP) method. The pose estimator distills from itself through the soft predictions of a previous batch, which are progressively refined with the model being updated. To ensure the difference between the teacher and the student, we propose a stepping strategy that widens the distillation distance when the learning rate degrades. Our main contribution is achieving efficient training and simple implementation at the same time, compared with previous pose distillation methods. Experimental results on two popular datasets validated the effectiveness of the proposed method.

6 Acknowledgments

This work has been supported by the project from the National Natural Science Foundation of China (92048203, 62376255)

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014) 6
2. Artacho, B., Savakis, A.: Unipose: Unified human pose estimation in single images and videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7035–7044 (2020) 1
3. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006) 2, 5
4. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 717–732. Springer (2016) 1
5. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 6
6. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017) 6
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016) 3
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 2, 5
9. Ke, L., Chang, M.C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018) 6
10. Li, Z., Ye, J., Song, M., Huang, Y., Pan, Z.: Online knowledge distillation for efficient pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11740–11750 (2021) 2, 4, 6, 8
11. 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: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 6
12. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016) 3, 6
13. Ning, G., Zhang, Z., He, Z.: Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Transactions on Multimedia **20**(5), 1246–1259 (2018). https://doi.org/10.1109/TMM.2017.2762010 6
14. Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. pp. 33–47. Springer (2014) 1
15. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019) 3, 5

16. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018) 6
17. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 648–656 (2015) 1
18. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. p. 1799–1807. NIPS'14, MIT Press, Cambridge, MA, USA (2014) 3, 6
19. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV) (2018) 1, 3, 6
20. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 1290–1299 (2017) 6
21. Zhang, F., Zhu, X., Ye, M.: Fast human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3517–3526 (2019) 2, 4, 6, 8