Orderly Dual-Teacher Knowledge Distillation for Lightweight Human Pose Estimation

Zhong-Qiu Zhao[1], Yao Gao[1], Yuchen Ge[1], Weidong Tian[1]
[1]Hefei University of Technology
{z.zhao,wdtian}@hfut.edu.cn,{2018110922,geyuchen}@mail.hfut.edu.cn

ABSTRACT

Although deep convolution neural networks (DCNN) have achieved excellent performance in human pose estimation, these networks often have a large number of parameters and computations, leading to the slow inference speed. For this issue, an effective solution is knowledge distillation, which transfers knowledge from a large pre-trained network (teacher) to a small network (student). However, there are some defects in the existing approaches: (I) Only a single teacher is adopted, neglecting the potential that a student can learn from multiple teachers. (II) The human segmentation mask can be regarded as additional prior information to restrict the location of keypoints, which is never utilized. (III) A student with a small number of parameters cannot fully imitate heatmaps provided by datasets and teachers. (IV) There exists noise in heatmaps generated by teachers, which causes model degradation. To overcome these defects, we propose an orderly dual-teacher knowledge distillation (ODKD) framework, which consists of two teachers with different capabilities. Specifically, the weaker one (primary teacher, PT) is used to teach keypoints information, the stronger one (senior teacher, ST) is utilized to transfer segmentation and keypoints information by adding the human segmentation mask. Taking dual-teacher together, an orderly learning strategy is proposed to promote knowledge absorbability. Moreover, we employ a binarization operation which further improves the learning ability of the student and reduces noise in heatmaps. Experimental results on COCO and OCHuman keypoints datasets show that our proposed ODKD can improve the performance of different lightweight models by a large margin, and HRNet-W16 equipped with ODKD achieves state-of-the-art performance for lightweight human pose estimation.

KEYWORDS

Human pose estimation, knowledge distillation, prior information, binarization operation

1 INTRODUCTION

Human pose estimation aims at locating the human keypoints in the input images. As a fundamental computer vision task, it has been applied in many areas such as human action recognition, virtual reality and smart surveillance.

With the development of DCNN [2, 6, 13, 19, 26, 28], human pose estimation has achieved significant improvements. On the challenging COCO benchmark [15], these networks consistently achieve top accuracy. However, the performance improvements always come with the cost of increasing the amount of parameters and computations, which leads to poor practicabilities on embedded devices such as smart phones and robots, like high memory requirements and slow inference speed. Li *et al.* [14] successfully constructed an in-home lower body rehabilitation system based on lightweight HRNet which makes the lightweight models receive more attention. Therefore, how to make a trade-off between performance and efficiency has become a crucial problem.

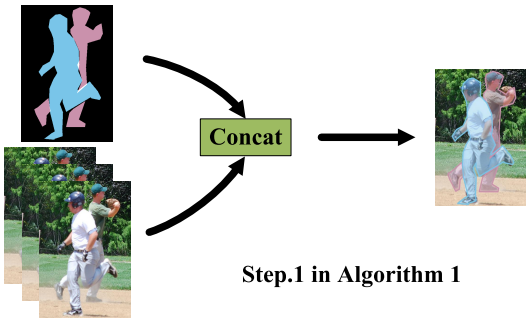


Figure 1: The process of introducing prior information.

There have been some attempts to address the above problem. On the basis of the large models [3, 26, 28], recent works [14, 20, 34] employed few stages or small backbones to achieve model compression. Although these models achieve a faster inference speed, their performance has a dramatic drop. Among the model compression approaches, knowledge distillation shows great superiority, which is to transfer knowledge of a large teacher model to a small student model. And a current work [32] has employed the student-teacher strategy to improve the performance and speed of the a student model. However, we observe that this method suffers from several problems: (I) Only a single teacher is used in the overall process, neglecting that multiple teachers can provide more privileged information for a student. (II) As a detection task, human pose estimation consists of two subtasks: the classification task (identifying the keypoints) and the regression task (locating the keypoints). It is difficult for large models to achieve high detection accuracy and more difficult for small lightweight models. Meanwhile, lightweight models are obtained by using some operations such as fewer stages and channels, which further causes inefficient heatmaps learning. (III) Ideally, in heatmaps generated by a human pose estimation model, there is only a peak with maximum response corresponding to the position of keypoints. However, multiple peaks usually appear in heatmaps, as pointed out in DARK [31]. Among them, only a peak is what we need to approximate and imitate. Redundant peaks can be deemed noise, which provides the wrong learning signal for a student model and causes the degeneration of a student model. Moreover, the existing human pose estimation models suffer from a common issue: in crowded scenarios, keypoints of one person are easily located on the body of another person, which severely affects the final performance. As available and extra prior information,

as shown in Figure 1, the human segmentation mask can provide valuable context cues to help restrict the position of keypoints, however, which is ignored.

To solve the above problems, we propose a new learning framework called orderly dual-teacher knowledge distillation (ODKD), which introduces two teachers with different capabilities. Specifically, the weaker one (primary teacher, PT) transfers keypoints information to a student, the stronger one (senior teacher, ST) teaches segmentation and keypoints information. These two teachers have slightly different inputs and network structures. Besides, as shown in TAKD [17], a student network performance degrades when the gap between a student and teacher is large. Considering that a similar situation also exists in our method, so we adopt the same strategy in TAKD where PT serves as a teacher assistant to bridge the gap between a student and ST.

Dual-teacher is employed in our proposed ODKD framework, a natural question to ask is: could we adopt more teachers to provide a student with more learning signals? In this paper, we mainly consider the following two aspects. Our proposed dual-teacher has different assignments where one teaches keypoints information, the other transfers segmentation and keypoints information. Adding additional teachers does not introduce new information. Besides, taking a dual-teacher is a compromise between performance and efficiency. The more teachers are, the better the performance may be but the longer the training time is. And we focus on exploring the knowledge distillation framework, rather than designing lightweight network structures. Therefore, we utilize two teachers in the final framework.

In a real-world scenario, when a small student network is asked to learn from two large teacher networks with different capabilities, a straightforward method is to learn from dual-teacher simultaneously. However, when facing two learning signals, the student is confused about which teacher it should learn from, and yielding a suboptimal result. From the perspective of human cognition, a more realistic and reasonable solution is multi-step learning. In theory, multi-step learning can stimulate more efficient learning of a student network. To this end, we propose an orderly learning strategy where the student learns from ST and PT successively so that knowledge from two aspects can be fully absorbed. To further improve the learning efficiency of a student network and reduce noise in heatmaps generated by teachers, we employ a binarization operation in [5] by which the student network only needs to classify each pixel in heatmaps as 0 or 1. And an appropriate binarization threshold can erase extra peaks, which avoids ineffective learning and model degeneration.

The purpose of adopting binarization operation in this paper is different from [5]. In [5], the binarization operation is utilized to convert the task so that Focal Loss can be used to address the class imbalance problem in human pose estimation. However, we employ a binarization operation to simplify the learning task and reduce noise.

To demonstrate the effectiveness and extensibility of our proposed ODKD framework, we conduct a series of experiments on two keypoints datasets, COCO [15] and OCHuman [33]. Experimental results show that ODKD can promote the existing lightweight human pose estimation models by a large margin.

The contributions of this paper are summarized as follows:

- In contrast to the existing works focusing on improving the performance of human pose estimation models, we pay more attention to the model efficiency. To this end, we propose an orderly dual-teacher knowledge distillation framework (ODKD) for human pose estimation, which integrates an orderly dual-teacher and the human segmentation mask. The proposed ODKD framework serves as a model-agnostic approach and can be applied to most of lightweight human pose estimation models.
- We adopt a binarization operation to convert the regression task to the classification task, and the task after binarization is simpler than the original task.
- We verify the effectiveness and extensibility of ODKD on different benchmark datasets, COCO and OCHuman with different baseline models. And HRNet-W16 equipped with ODKD achieves state-of-the-art performance for lightweight human pose estimation.

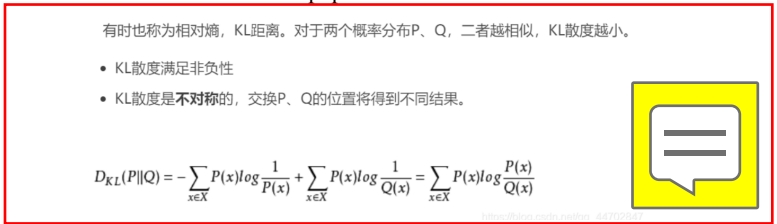
2 RELATED WORK

2.1 Lightweight Human Pose Estimation

There have been some works to compress the human pose estimation models. Based on OpenPose [3], Daniil *et al.* [20] proposed a lightweight OpenPose network where the heavy computational backbone VGG [24] was replaced by the simple and effective MobileNet [11], and 7×7 convolutions were replaced by lots of 3×3 convolutions. Although its inference speed becomes faster, there is a significant performance gap between the model and the current mainstream models. Umer *et al.* [23] constructed an efficient convolutional network to accelerate inference without conducting quantitative experiments on model efficiency. Bulat *et al.* [1] employed the neural network binarization to achieve model compression. Zhang *et al.* [34] proposed a lightweight pose network (LPN), which was equipped with the depthwise separable convolution and iterative training strategy. However, this iterative training strategy takes more than triple the training time of the original training strategy. Li *et al.* [14] proposed a lightweight HRNet which integrated the attention mechanism with Efficient Spatial Pyramid (ESP) [16].

2.2 Knowledge Distillation

Knowledge distillation has been successfully applied to many computer vision tasks, such as image classification [12, 17, 25, 27, 29], object detection [4, 8], pose estimation [32, 35], etc. Chen *et al.* [4] employed knowledge distillation to learn compact object detection networks and proposed several loss functions to improve the efficiency of knowledge transfer. Dai *et al.* [8] proposed a general instance distillation for object detection where feature-based, relation-based and response-based knowledge was considered. Zhao *et al.* [35] employed a knowledge framework to solve the occlusion problem in human pose estimation. Zhang *et al.* [32] expanded the lightweight human pose estimation network by introducing knowledge distillation, but the approach only employed a single teacher, neglecting that multiple teachers can provide more valuable information. Mirzadeh *et al.* [17] introduced a distillation framework called Teacher Assistant Knowledge Distillation (TAKD), where an intermediate-sized network (teacher assistant) was adopted to



bridge the gap between the student and the teacher. Song *et al.* [25] proposed a densely guided knowledge distillation (DGKD) to eliminate the error avalanche problem in TAKD. The main difference between our method and TAKD and DGKD can be summarized as follows: (I) Our method employs dual-teacher to teach different information, while TAKD and DGKD apply multiple teachers to teach the same information. (II) Although our method is inspired by DGKD, there is some difference between them. When distilling the student model, our method employs an orderly learning strategy, while DGKD utilizes multiple teachers to teach a student simultaneously. Crucially, our method can be seen as an extension of existing knowledge distillation methods.

3 APPROACH

3.1 Background

The key to knowledge distillation is to let a small network (student) imitate not only the output of a large network (teacher), but also true labels of datasets. Let l_s and l_t be the logits of the student and teacher, respectively. T is a temperature parameter to soften the output of the student and teacher. $y_s = softmax(l_s/T)$ and $y_t = softmax(l_t/T)$ are the soften outputs of the student and teacher, respectively. To encourage the student to mimic the output of the teacher, a KL-divergence loss L_{KD} can be minimized as follows:

$$L_{KD} = T^2 KL(y_s, y_t) \tag{1}$$

To minimize the gap between the output of the student model softmax(l_s) and true labels l of datasets, the cross-entropy loss L_{CE} can be penalized as follows:

$$L_{CE} = F(softmax(l_s), l) \tag{2}$$

Finally, the overall loss function can be denoted by adding a balance factor α as follows:

$$L = (1 - \alpha)L_{CE} + \alpha L_{KD} \tag{3}$$

3.2 ODKD Framework

As illustrated in Figure 2, our proposed orderly dual-teacher knowledge distillation (ODKD) framework consists of two large teacher networks with different capabilities: primary teacher (PT) and senior teacher (ST). The final target is to transfer knowledge of PT and ST to a lightweight student network. We adopt a binarization operation and two different loss functions. In this section, we firstly elaborate on the implementation process of an orderly dual-teacher. Secondly, we describe the details of a binarization operation. Finally, we give loss functions used in the overall process.

Orderly dual-teacher. There are three differences between PT and ST. The first is the input of networks. The input of PT is a three-channel RGB image, while the input of ST is generated by concatenating a three-channel RGB image and one-channel human segmentation mask. The second is slightly different network structures. PT is a common human pose estimation model, such as SimpleBaseline [28], while ST is the variant of PT where we add an 1×1 convolution at the head of ST to transform the number of channels from 4 to 3. The third is different capabilities. PT is used to teach keypoints information, but ST is used to transfer segmentation and keypoints information. There is some difference between a lightweight student and PT. For example, a student model

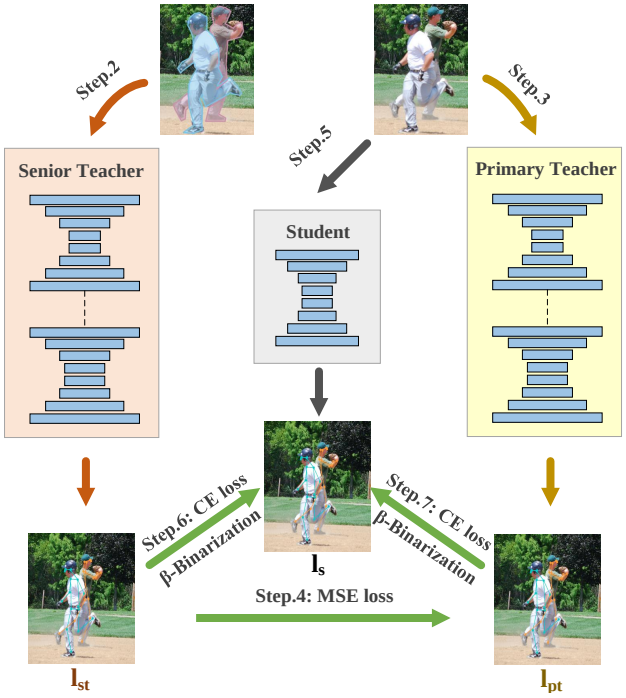


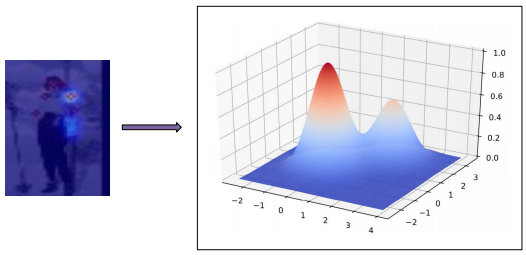
Figure 2: Illustration of our method. We adopt two loss functions: MSE loss and cross-entropy loss. β-Binarization denotes the binarization operation with a factor β.

employs fewer channels or stages than PT, which leads to a large performance gap between them. With the guidance of the human segmentation mask, ST is more powerful than PT, which further widens the gap between a student and ST. To solve the problem, as shown in Step 4 of Figure 2, we add a path where PT serves as a teacher assistant to bridge the gap between a student and ST. During training, we firstly pre-train ST by imitating true labels of datasets. Then PT is obtained by mimicking labels provided by datasets and ST. After these, an orderly learning strategy is utilized where a student learns from ST and PT successively, as shown in Steps 6 and 7 of Figure 2. At test time, only a student is employed.

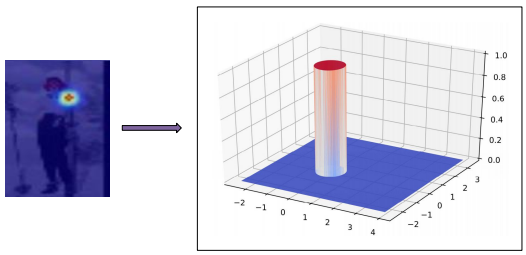
Binarization operation. The binarization operation aims to simplify the learning task and reduce noise in heatmaps generated by teachers. As shown in Figure 3(a), the heatmap predicted by teacher models almost exhibits a 2D Gaussian distribution structure where each pixel of the heatmap ranges from 0 to 1. It is very difficult for a lightweight student model to approximate the distribution. A straightforward solution is to convert the difficult task to a simple task. Moreover, multiple peaks appear in heatmaps generated by teachers and redundant peaks can be regarded as noise, which provides a worthless learning signal for a student. To solve the above problems, as shown in Figure 2, we employ a binarization operation during obtaining a lightweight student model, which can be denoted as follows with a threshold β:

$$C(y) = \begin{cases} 1 & if\ H(y) > \beta, \\ 0 & if\ H(y) \le \beta, \end{cases} \tag{4}$$

where $H(y)$ is the value of the heatmap at location y and $C(y)$ is the class of pixel at location y. For the ground-truth heatmaps,



(a) Heatmap predicted by teachers



(b) Heatmap after a binarization operation

Figure 3: Illustration of a binarization operation. (a) Predicted Heatmap. (b) Heatmap representation after a binarization operation.

we empirically set β to 0.6. For the heatmaps from teachers, the threshold is chosen by the ablation experiments. After binarization, the result is illustrated in Figure 3(b). A student only needs to make a simple binary classification for each pixel and an appropriate threshold can eliminate redundant peaks. When training ST and PT, we do not employ a binarization operation as the number of layers of these two models is enough to approximate the data distribution.

Loss functions. Whether to use a binarization operation affects which loss function to use. When training ST and PT, a binarization operation is not employed, and therefore the loss function used is the conventional MSE loss, which can be denoted as follows:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{l}^i - l^i)^2, \tag{5}$$

where $\hat{l}^i$ and l^i specify the predicted heatmap and ground-truth heatmap for the i-th joint, respectively. In the process of obtaining a lightweight student model, a binarization operation is utilized, and the cross-entropy loss is penalized to minimize the gap between the predicted value $q = Sigmoid(l_s)$ and the label value p:

$$\begin{aligned} L_C &= -(plogq + (1-p)log(1-q)) \\ &= \begin{cases} -logq & if\ p = 1, \\ -log(1-q) & if\ p = 0. \end{cases} \end{aligned} \tag{6}$$

Taking Eq.(3), (5), and (6) together, we can obtain the loss functions used in the overall process as follows:

$$L_{ST} = L_{MSE} \tag{7}$$

$$\begin{aligned} L_{PT} = L_{ST\to PT} &= (1-\alpha_0)L_{CE_{PT}} + \alpha_0 L_{KD_{ST\to PT}} \\ &= (1-\alpha_0)L_{MSE_{PT}} + \alpha_0 L_{MSE_{ST\to PT}}, \end{aligned} \tag{8}$$

Algorithm 1 : ODKD student training

Input: preprocessed image x and human segmentation mask m, label l, pre-trained senior teacher ST, primary teacher PT, student S, the number of iterations n, the number of epochs N

Output: distilled student S.

```
for i = 1 to N do
  for j = 1 to n do
    Step.1 concatenate x and m to obtain y;
    Step.2 feed y to ST, to obtain the senior teacher logits l_st;
    Step.3 feed x to PT, to obtain the primary teacher logits l_pt;
    Step.4 update PT based on Eq.(8);
    Step.5 feed x to S, to obtain the student logits l_s;
    Step.6 update S based on Eq.(9);
    Step.7 update S based on Eq.(10).
  end for
end for
```

$$\begin{aligned} L_{S_1} = L_{ST\to S} &= (1-\alpha_1)L_{CE_S} + \alpha_1 L_{KD_{ST\to S}} \\ &= (1-\alpha_1)L_{C_S} + \alpha_1 L_{C_{ST\to S}}, \end{aligned} \tag{9}$$

$$\begin{aligned} L_{S_2} = L_{PT\to S} &= (1-\alpha_2)L_{CE_S} + \alpha_2 L_{KD_{PT\to S}} \\ &= (1-\alpha_2)L_{C_S} + \alpha_2 L_{C_{PT\to S}}, \end{aligned} \tag{10}$$

where the right arrow at the subscript indicates the teaching direction, and α_0, α_1 and α_2 are balance factors. Here we set them to 0.5 as demonstrated in FPD [32]. Finally, the overall process can be summarized as Algorithm 1.

4 EXPERIMENT

4.1 Experiment Setup

Dataset. We employ two human pose estimation datasets, COCO and OCHuman. COCO dataset contains over 200K images and 250K person instances labeled with 17 keypoints. The images are extracted from real scenes. We train our models only on the train2017 set, equipped with 57K images and 150K person instances, and evaluate our method on the val2017 set and test-dev2017 set, consisting of 5K images and 20K images, respectively.

OCHuman dataset is also collected from real scenes. Different from COCO dataset, it is a more challenging dataset, where each human instance is heavily occluded by one or several others and the postures of the human bodies are more complex. The purpose of designing this dataset is to use general datasets such as COCO as a training set, test the robustness of models to occlusion using OCHuman. Therefore, this dataset has no training set, but only a validation set and test set. The validation set and test set have 4731 images and 8110 person instances in total.

Evaluation Metric. For the two datasets, we employ the same evaluation metric based on Object Keypoint Similarity (OKS). OKS can be calculated by:

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{11}$$

where, d_i is the Euclidean distance between each ground truth keypoint and corresponding detected keypoint, v_i is the visibility flag of the ground truth, s is the object scale, and k_i is a per-keypoint constant that controls falloff. We report standard average precision and recall: AP (the mean of AP scores at OKS = 0.50, 0.55,. . ., 0.90, 0.95), AP^{50} (AP at OKS = 0.50), AP^{75}, AP^M for medium objects, AP^L for large objects and AR (the mean of AR scores at OKS = 0.50, 0.55,. . ., 0.90, 0.95).

Training. Our models are implemented on two NVIDIA 2080Ti GPUs. We extend the human detection boxes to a fixed ratio, namely height: width = 4 : 3, and then crop the boxes from images. Finally, we resize the cropped images to a fixed size, for example, 256×192. In the experiments, we choose three combinations of ST, PT and student models. One combination is that SimpleBaseline-ResNet50 [28] with the human segmentation mask, SimpleBaseline-ResNet50 and LPN-ResNet50 [34] are chosen as ST, PT and student, respectively. One combination is an 8-stage hourglass [32] with the human segmentation mask, 8-stage hourglass and 4-stage hourglass. Another combination is HRNet-W32 [26] with the human segmentation mask, HRNet-W32 and HRNet-W16. We reproduce LPN without any tricks, including the attention mechanism, the iterative training strategy and β-Soft-Argmax. Other settings are the same as the original work. For SimpleBaseline and Hourglass network, we adopt the same training strategies as in the original works. Following the structure of HRNet, we design a small network HRNet-W16 by reducing the number of basic channels to 16. The total epochs are set as 150, and the learning rate is dropped at the 120th and 140th epochs.

Testing. The top-down pipeline is adopted that first locates the human body by the person detectors and then applies the pose estimation models to acquire the detection results. For a fair comparison, we adopt the same person detectors provided by HRNet [26] both for COCO validation and test-dev set. The human detection AP is 56.4 and 60.9 respectively. For OCHuman test set, we use ground-truth detection boxes. Following the common practice[6, 19, 26, 28], we compute the heatmap by averaging the heatmaps of the original and flipped images. The final keypoints are obtained by adjusting a quarter offset in the direction from the highest response to the second highest response.

4.2 Component Ablation Studies

In this subsection, we conduct the ablation experiments on the COCO validation set to verify the effectiveness of our proposed components. We choose SimpleBaseline as the teacher model and LPN as the student model. By default, the input size 256×192 and the ResNet-50 backbone are used because they are less computational.

The studies on distillation paths. We explore the performance of different distillation paths on the COCO validation set. As shown in Table 1, we divide all the experiments into four groups. The first group is our baseline model, which achieves 64.5 AP. The second group belongs to single teacher knowledge distillation. The third group belongs to dual-teacher knowledge distillation, where there is no path between ST and PT. The fourth group is dual-teacher knowledge distillation, where PT serves as a teacher assistant to bridge the gap between ST and S. We firstly compare these methods within each group. In group 2, Method (b) obtains less performance improvement than (c), as the output of ST is more abstract than that of PT, which is a more difficult learning task for a student. In group 3, Method (d) is a situation where dual-teacher teaches a student simultaneously, while Method (e) and (f) belong to a circumstance where dual-teacher teaches a student in multiple steps. Method (d) and (f) achieve 65.0 AP, which is higher than (e). The reason behind this is that in Method (d) and (f), segmentation information can be firstly used to help restrict the location range of keypoints, and then keypoints information is utilized to optimize the location of keypoints, which produces more accurate heatmaps. While in Method (e), segmentation information is learned in the second step, which leads to inadequate utilization of prior information. In group 4, Method (h) employs dual-teacher to guide a student simultaneously, while Method (i) and (j) utilize an orderly dual-teacher knowledge distillation strategy. Method (i) and (j) receive more improvement than (h), which proves the effectiveness of an orderly learning strategy. Method (j) gets higher AP than (i), which further demonstrates the superiority of firstly learning segmentation information. Comparisons of methods in different groups are shown as follows. Compared with group 1, all of the other groups achieve performance improvements, which shows the advantage of knowledge distillation. Group 2 and 3 achieve similar performance. By adding a path on the basis of group 3, group 4 can be obtained, which receives a better performance than the original, which demonstrates that it is necessary to introduce PT as a teacher assistant to narrow the gap between ST and S. Compared to Method (c), Method (j) obtains 0.2 AP improvement, which is not obvious. The main reason is the insufficient learning ability of a student, rather than an orderly dual-teacher learning strategy.

Table 1: The ablation study on the architecture of ODKD on the COCO validation set. We adopt MSE Loss. → indicates the teaching direction.

Group	Method	AP
1	(a) S (LPN, baseline)	64.5
2	(b) ST→S	64.9
	(c) PT→S	65.0
3	(d) ST, PT→S	65.0
	(e) PT→S, ST→S	64.7
	(f) ST→S, PT→S	65.0
4	(g) ST→PT→S	64.9
	(h) ST→PT. ST, PT→S	64.9
	(i) ST→PT. PT→S, ST→S	65.0
	(j) ST→PT. ST→S, PT→S (ODKD)	**65.2**

Table 2: Comparisons on the binarization thresholds and GPU memory usage. The batch size is set as 24.

Models	LPN	ODKD				
β	-	0.2	0.3	0.4	0.5	0.6
AP	65.2	65.7	**65.9**	65.7	65.3	65.1
GPU Memory Usage (MB)	9286	**6694**				

Table 3: Comparisons on the COCO validation set. The result of 8-stage Hourglass is cited from [19]. 8-stage Hourglass* and 4-stage Hourglass*: models reproduced by ourselves using the COCO dataset.

Method	Backbone	Pretrain	Input size	#Params	FLOPs	FPS	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Large networks												
8-stage Hourglass [19]	8-stage Hourglass	N	256×192	25.1M	14.3G	-	66.9	-	-	-	-	-
HRNet-W32 [26]	HRNet-W32	N	256×192	28.5M	7.1G	140	73.4	89.5	80.7	70.2	80.1	78.9
8-stage Hourglass* [32]	8-stage Hourglass	N	256×192	25.6M	21.3G	49	73.7	89.3	80.7	70.4	80.4	79.2
CPN [6]	ResNet-50	Y	256×192	27.0M	6.2G	77	69.4	-	-	-	-	-
SimpleBaseline [28]	ResNet-50	Y	256×192	34.0M	8.9G	187	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [28]	ResNet-101	Y	256×192	53.0M	12.4G	155	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [28]	ResNet-152	Y	256×192	68.6M	15.7G	124	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32 [26]	HRNet-W32	Y	256×192	28.5M	7.1G	140	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [26]	HRNet-W48	Y	256×192	63.6M	14.6G	103	75.1	90.6	82.2	71.5	81.8	80.4
Small networks												
Lite-HRNet [30]	Lite-HRNet-18	N	256×192	**1.1M**	**0.2G**	-	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet [30]	Lite-HRNet-30	N	256×192	**1.8M**	**0.3G**	-	67.2	88.0	75.0	64.3	73.1	73.3
LPN50 [34]	ResNet-50	N	256×192	**2.7M**	**1.1G**	**243**	64.5	86.3	71.8	61.1	71.1	70.7
+ODKD	ResNet-50	N	256×192	**2.7M**	**1.1G**	**243**	**65.9(+1.4)**	**86.9**	**73.1**	**62.5**	**72.8**	**72.0**
4-stage Hourglass* [32]	4-stage Hourglass	N	256×192	**3.3M**	**3.0G**	**158**	68.3	87.1	75.4	65.3	74.3	74.1
+ODKD	4-stage Hourglass	N	256×192	**3.3M**	**3.0G**	**158**	**69.3(+1.0)**	**87.4**	**76.6**	**66.2**	**75.8**	**75.1**
HRNet-W16 [26]	HRNet-W16	N	256×192	**7.5M**	**2.6G**	**163**	68.4	88.3	76.7	65.2	74.7	74.7
+ODKD	HRNet-W16	N	256×192	**7.5M**	**2.6G**	**163**	**71.7(+3.3)**	**89.3**	**79.1**	**68.7**	**78.0**	**77.5**

Among all methods, Method (j) achieves the best 65.2 AP. Therefore, this setting is selected finally.

The binarization thresholds of heatmaps from teachers. We compare the effects of the heatmaps with different binarization thresholds on the COCO validation set. As shown in Table 2, with increasing β, the performance first increases and then decreases. The binarization thresholds represent the degree of reducing noise. The larger it is, the stronger the degree of reducing noise is. When β is set to 0.3, lots of noise is removed and a model achieves the best 65.9 AP. When it is large enough such as 0.6, part of the learning signal in heatmaps will be eliminated, which leads to model degradation. Finally, β is set as 0.3. A binarization operation can reduce GPU memory usage, which is friendly to resource-limited devices. Moreover, a binarization operation can effectively improve the model performance as illustrated in Figure 4.

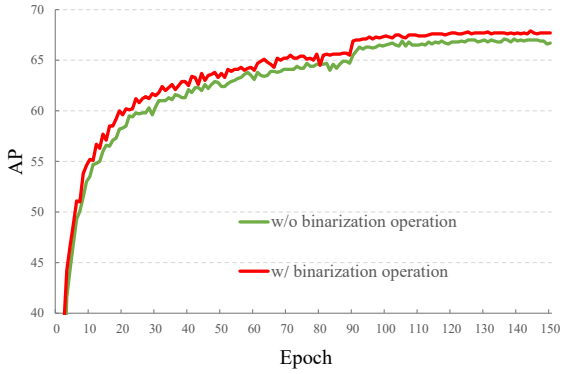


Figure 4: Illustration of performance with or without a binarization operation.

4.3 Experimental Results

Results on the COCO validation set. We report the results of our method and other state-of-the-art methods in Table 3. (I) When ODKD is applied to lightweight models such as LPN50, 4-stage Hourglass and HRNet-W16, performance can be consistently improved, which demonstrates the superiority of the proposed ODKD. However, the improvements on different baseline models are discrepant. The reason for this is that the gaps between a student and teacher within each combination are different. The difference between LPN50 and SimpleBaseline [28] is that LPN50 employs a depth-wise separable convolution while SimpleBaseline uses an ordinary convolution. There are two differences between 4-stage Hourglass and 8-stage Hourglass. One is different numbers of stages, the other is different numbers of channels. The only difference between HRNet-W16 and HRNet-W32 is the number of channels. The model similarity between HRNet-W16 and HRNet-W32 is higher than other combinations, which yields higher performance improvement. (II) Compared to 8-stage Hourglass and CPN [6], LPN50 and 4-stage Hourglass based on ODKD achieve comparable performance, while the parameters and calculation are much less than them. (III) HRNet-W16 trained with ODKD obtains better performance than SimpleBaseline with ResNet-50 and ResNet-101. And we can find that HRNet-W16 with 2.6GFLOPs runs slower than SimpleBaseline-ResNet50 with 8.9GFLOPs because there are lots of parallel convolutions in HRNet and PyTorch framework is not friendly to parallel convolutions. (IV) Although HRNet-W32 and HRNet-W48 achieve top accuracy, inference time is longer than

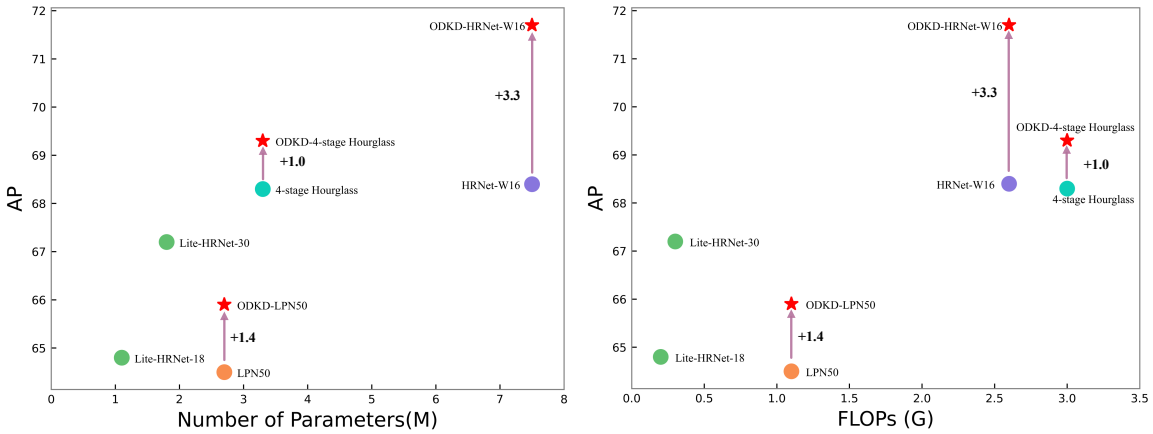


Figure 5: Illustration of the complexity and accuracy comparison on the COCO validation set.

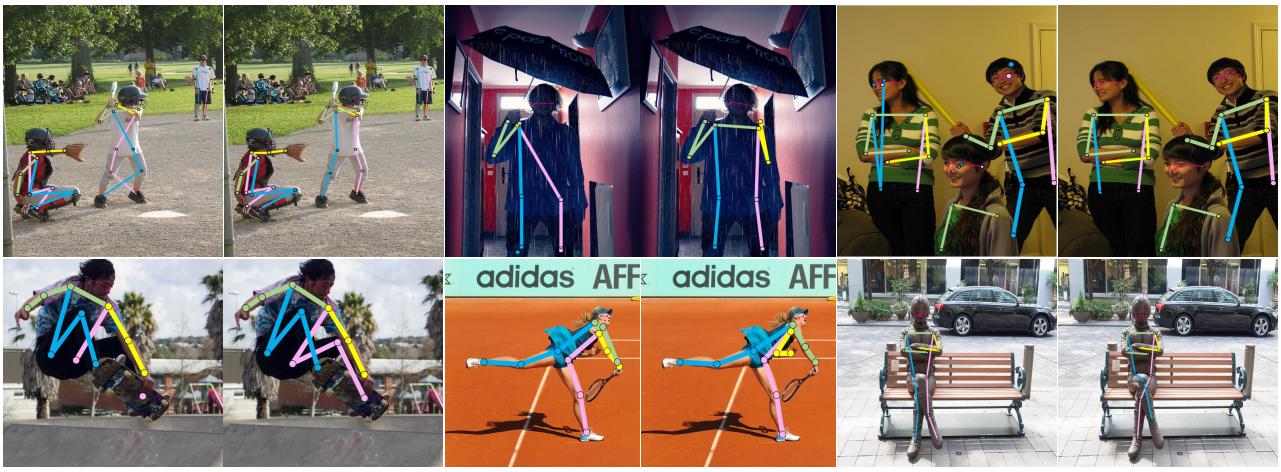
Table 4: Comparisons on the COCO test-dev set. 4-stage Hourglass* and 8-stage Hourglass*: models reproduced by ourselves using the COCO dataset.

Method	Backbone	Input size	#Params	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Large networks										
OpenPose [3]	-	-	-	-	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [18]	-	-	-	-	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab[21]	-	-	-	-	68.7	89.0	75.4	64.1	75.5	75.4
HigherHRNet[7] (multi-scale test)	HRNet-W48	640×640	63.8M	154.3G	70.5	89.3	77.2	66.6	75.8	74.9
Mask-RCNN[10]	ResNet-50-FPN	-	-	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI[22]	ResNet-101	353×257	42.6M	57.0G	64.9	85.5	71.3	62.3	70.0	69.7
CPN [6]	ResNet-50	256×192	27.0M	6.2G	68.6	89.5	76.6	65.6	74.2	75.6
RMPE [9]	PyraNet	320×256	28.1M	26.7G	72.3	89.2	79.1	68.0	78.6	-
SimpleBaseline [28]	ResNet-50	256×192	34.0M	8.9G	70.0	90.9	77.9	66.8	75.8	75.6
HRNet-W32 [26]	HRNet-W32	256×192	28.5M	7.1G	73.5	92.2	81.9	70.2	79.2	79.0
HRNet-W48 [26]	HRNet-W48	256×192	63.6M	14.6G	74.2	92.4	82.4	70.9	79.7	79.5
8-stage Hourglass* [32]	8-stage Hourglass	256×192	25.6M	21.3G	73.2	91.3	81.1	70.2	79.0	78.7
Small networks										
Lite-HRNet [30]	Lite-HRNet-18	384×288	**1.1M**	**0.45G**	66.9	89.4	74.4	64.0	72.2	72.6
Lite-HRNet [30]	Lite-HRNet-30	384×288	**1.8M**	**0.7G**	69.7	90.7	77.5	66.9	75.0	75.4
LPN50 [34]	ResNet-50	256×192	**2.7M**	**1.1G**	64.2	88.6	71.2	61.0	69.8	70.1
+ODKD	ResNet-50	256×192	**2.7M**	**1.1G**	**65.5(+1.3)**	**89.2**	**72.8**	**62.4**	**71.1**	**71.4**
4-stage Hourglass* [32]	4-stage Hourglass	256×192	**3.3M**	**3.0G**	67.8	89.1	75.4	64.9	73.2	73.4
+ODKD	4-stage Hourglass	256×192	**3.3M**	**3.0G**	**69.1(+1.3)**	**89.9**	**76.7**	**66.0**	**74.7**	**74.7**
HRNet-W16 [26]	HRNet-W16	256×192	**7.5M**	**2.6G**	67.6	90.2	76.1	64.7	73.0	73.7
+ODKD	HRNet-W16	256×192	**7.5M**	**2.6G**	**71.0(+3.4)**	**91.1**	**79.5**	**67.9**	**76.6**	**76.6**

our models. (V) Compared to the latest Lite-HRNet[1], our models based on ODKD achieve a better balance between accuracy and computational complexity, as shown in Figure 5. We think that the proposed ODKD is a general framework and can be applied to Lite-HRNet to further improve the performance of the model. Some results generated by baseline models and our models are illustrated in Figure 6. We can see that our models can work well in challenging situations.

Results on the COCO test-dev set. Table 4 shows the results of our method and other state-of-the-art methods. Our proposed ODKD can promote LPN50 and 4-stage Hourglass by 1.3 AP. For HRNet-W16, the improvement is 3.4 AP. Compared to the bottom-up approaches, our models achieve acceptable results with fewer parameters and FLOPs. HRNet-W16 based on ODKD is significantly

[1]Available from https://github.com/HRNet/Lite-HRNet. Because this paper and the corresponding code have just been published for few days, the experiments of adding ODKD to Lite-HRNet are not finished and the results are not displayed.



LPN LPN (Ours) Hourglass Hourglass (Ours) HRNet-W16 HRNet-W16 (Ours)

Figure 6: Qualitative results on COCO validation set.

Table 5: Comparisons on the OCHuman test set. 4-stage Hourglass* and 8-stage Hourglass*: models reproduced by ourselves using the COCO dataset. There is no medium person instances, so AP^M denotes as -.

Method	Backbone	Input size	#Params	FLOPs	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
Large networks										
CPN [6]	ResNet-50	256 × 192	27.0M	6.2G	52.8	75.8	55.2	-	52.8	61.6
SimpleBaseline [28]	ResNet-50	256 × 192	34.0M	8.9G	55.8	73.9	61.3	-	55.8	60.4
SimpleBaseline [28]	ResNet-101	256 × 192	53.0M	12.4G	60.0	76.0	66.1	-	60.0	63.9
SimpleBaseline [28]	ResNet-152	256 × 192	68.6M	15.7G	61.9	77.0	67.8	-	61.9	66.1
HRNet-W32 [26]	HRNet-W32	256 × 192	28.5M	7.1G	63.0	79.2	68.5	-	63.0	66.9
HRNet-W48 [26]	HRNet-W48	256 × 192	63.6M	14.6G	64.6	79.6	70.7	-	64.6	68.1
8-stage Hourglass* [32]	8-stage Hourglass	256 × 192	25.6M	21.3G	64.7	79.6	69.9	-	64.7	68.3
Small networks										
LPN50 [34]	ResNet-50	256 × 192	**2.7M**	**1.1G**	49.6	**73.4**	54.2	-	49.6	54.9
+ODKD	ResNet-50	256 × 192	**2.7M**	**1.1G**	**50.3(+0.7)**	72.1	**54.6**	-	**50.3**	**55.9**
4-stage Hourglass* [32]	4-stage Hourglass	256 × 192	**3.3M**	**3.0G**	56.3	75.0	62.6	-	56.3	60.9
+ODKD	4-stage Hourglass	256 × 192	**3.3M**	**3.0G**	**58.6(+2.3)**	**77.0**	**64.6**	-	**58.6**	**63.0**
HRNet-W16 [26]	HRNet-W16	256 × 192	**7.5M**	**2.6G**	56.9	76.9	63.3	-	56.9	61.4
+ODKD	HRNet-W16	256 × 192	**7.5M**	**2.6G**	**60.5(+3.6)**	**79.3**	**66.7**	-	**60.5**	**64.9**

better than CPN [6] and SimpleBaseline [28]. Compared to HRNet-W32 and HRNet-W48, HRNet-W16 trained with ODKD obtains close performance without pre-training. Compared to Lite-HRNet, our models achieve better performance with similar parameters and smaller input size. Moreover, HRNet-W16 equipped with ODKD sets a new state-of-the-art for lightweight human pose estimation.

Results on the OCHuman test set. We show the results of our methods and other state-of-the-art methods in Table 5. We use the models trained on COCO dataset to evaluate on OCHuman test set. As OCHuman dataset is more challenging than COCO dataset, the performance of all methods drops sharply. Different baseline models can benefit from ODKD, which proves the effectiveness and extensibility of ODKD. Moreover, our models based on ODKD use fewer parameters achieve comparable accuracy with large models, which is more suitable for mobile devices.

5 CONCLUSION

In this paper, we propose an orderly dual-teacher knowledge distillation (ODKD) framework for human pose estimation. Dual-teacher is introduced, where one is used to teach keypoints information to a student, the other is utilized to transfer segmentation and keypoints information. To improve the learning ability of a student, an orderly learning strategy is adopted where a student learns from dual-teacher successively. Moreover, a binarization operation is

employed to stimulate more efficient learning of the a student and reduce noise in heatmaps generated by teachers. The experimental results on COCO and OCHuman keypoints dataset demonstrate the effectiveness and extensibility of our proposed ODKD.

REFERENCES

[1] Adrian Bulat and Georgios Tzimiropoulos. 2017. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*. 3706–3714.

[2] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. 2020. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*. 455–472.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 742–751.

[5] Xiao Chen and Genke Yang. 2018. Multi-person pose estimation with LIMB detection heatmaps. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 4078–4082.

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.

[7] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5386–5395.

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. 2021. General Instance Distillation for Object Detection. *arXiv preprint arXiv:2103.02340* (2021).

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2334–2343.

[10] K He, G Gkioxari, and PY Girshick Dollár. 2017. R., 2017. Mask RCNN. In *2017 IEEE International Conference on Computer Vision*. 1440–1448.

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[12] Nikos Komodakis and Sergey Zagoruyko. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

[13] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. 2019. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148* (2019).

[14] Ying Li, Chenxi Wang, Yu Cao, Benyuan Liu, Joanna Tan, and Yan Luo. 2020. Human pose estimation based in-home lower body rehabilitation system. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[16] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. 2018. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*. 552–568.

[17] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5191–5198.

[18] Alejandro Newell, Zhiao Huang, and Jia Deng. 2016. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424* (2016).

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*. Springer, 483–499.

[20] Daniil Osokin. 2018. Real-time 2d multi-person pose estimation on CPU: Lightweight OpenPose. *arXiv preprint arXiv:1811.12004* (2018).

[21] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 269–286.

[22] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. 2017. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4903–4911.

[23] Umer Rafi, Bastian Leibe, Juergen Gall, and Ilya Kostrikov. 2016. An Efficient Convolutional Network for Human Pose Estimation.. In *BMVC*, Vol. 1. 2.

[24] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[25] Wonchul Son, Jaemin Na, and Wonjun Hwang. 2020. Densely Guided Knowledge Distillation using Multiple Teacher Assistants. *arXiv preprint arXiv:2009.08825* (2020).

[26] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5693–5703.

[27] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2018. KDGAN: Knowledge Distillation with Generative Adversarial Networks.. In *NeurIPS*. 783–794.

[28] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 466–481.

[29] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4133–4141.

[30] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. 2021. Lite-HRNet: A Lightweight High-Resolution Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. 2020. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7093–7102.

[32] Feng Zhang, Xiatian Zhu, and Mao Ye. 2019. Fast human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3517–3526.

[33] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. 2019. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 889–898.

[34] Zhe Zhang, Jie Tang, and Gangshan Wu. 2019. Simple and Lightweight Human Pose Estimation. *arXiv preprint arXiv:1911.10346* (2019).

[35] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.