Online Knowledge Distillation for Efficient Pose Estimation

Zheng Li [1], Jingwen Ye [2], Mingli Song [2], Ying Huang [1], Zhigeng Pan [1*]
[1] Hangzhou Normal University, [2] Zhejiang University
lizheng1@stu.hznu.edu.cn, {yejingwen,brooksong}@zju.edu.cn, {yw52,zgpan}@hznu.edu.cn

Abstract

*Existing state-of-the-art human pose estimation methods require heavy computational resources for accurate predictions. One promising technique to obtain an accurate yet lightweight pose estimator is knowledge distillation, which distills the pose knowledge from a powerful teacher model to a less-parameterized student model. However, existing pose distillation works rely on a heavy pre-trained estimator to perform knowledge transfer and require a complex two-stage learning procedure. In this work, we investigate a novel **O**nline **K**nowledge **D**istillation framework by distilling **H**uman **P**ose structure knowledge in a one-stage manner to guarantee the distillation efficiency, termed **OKDHP**. Specifically, OKDHP trains a single multi-branch network and acquires the predicted heatmaps from each, which are then assembled by a Feature Aggregation Unit (FAU) as the target heatmaps to teach each branch in reverse. Instead of simply averaging the heatmaps, FAU which consists of multiple parallel transformations with different receptive fields, leverages the multi-scale information, thus obtains target heatmaps with higher-quality. Specifically, the pixel-wise Kullback-Leibler (KL) divergence is utilized to minimize the discrepancy between the target heatmaps and the predicted ones, which enables the student network to learn the implicit keypoint relationship. Besides, an unbalanced OKDHP scheme is introduced to customize the student networks with different compression rates. The effectiveness of our approach is demonstrated by extensive experiments on two common benchmark datasets, MPII and COCO.*

1. Introduction

Human pose estimation aims to recognize and localize all the human anatomical keypoints in a single RGB image. It's a fundamental technique for high-level vision tasks, such as action recognition [11], virtual reality [44] and human-computer interaction. Since the invention of DeepPose [55], deep neural networks have been the dominant solution for human pose estimation, based on which, the approaches [57, 61, 52] focus on exploiting richer representations with a sequential architecture and achieve state-of-the-art performance. However, the gains of such deep learning based approaches often come with a cost of training and deploying the over-parameterized models, which limits the deployment in resource-intensive mobile devices. To reduce the computation cost and enhance the model efficiency, many efforts have been devoted to directly designing lightweight and real-time networks, e.g., PAF [4], VNect [36] and MultiPoseNet [27].

As another powerful tool to achieve a good trade-off between speed and accuracy, knowledge distillation [19] follows the teacher-student paradigm. Traditional distillation utilizes a two-stage scheme that starts with a cumbersome pre-trained teacher model, then distills the knowledge to a compact student model. In the field of pose estimation, recent works [66, 22] adopt a traditional offline distillation scheme which distills the knowledge from a large pre-trained pose estimator (teacher) to a lightweight pose estimator (student) as shown in Fig. 1(a). However, training such a heavy teacher model is time-consuming and a

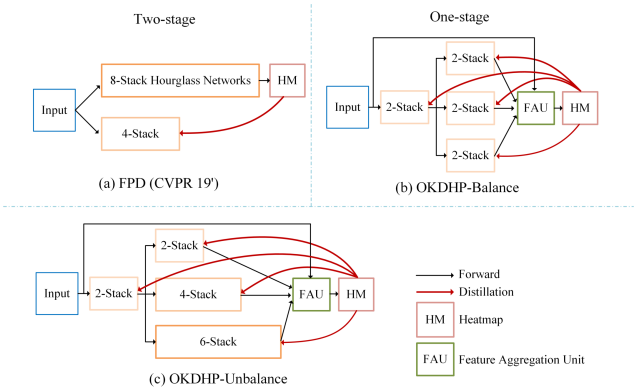

Figure 1. To obtain an efficient 4-Stack network, (a) FPD [66] adopts a two-stage distillation scheme from the static pre-trained 8-Stack network. The proposed OKDHP distills the pose structural knowledge with both (b) Balance and (c) Unbalance architectures in one stage. The teacher is established online with the FAU.

*Corresponding author

high-capacity model is not always available. Thus, online counterparts [67, 68] are proposed to simplify the distillation process to one stage, reducing the demand for the pre-trained teacher model. In ONE [68], a strong teacher model is established on-the-fly and all students share the same target distribution by averaging the predictions of all branches with learnable weights. Prior impressive works are mostly devoted to classification tasks, which neglect the valuable structural knowledge in the pixel-level tasks. Thus, our work focuses on the more challenging pixel-level tasks and proposes the first online pose distillation framework.

Existing pixel-level distillation works [66, 22] use mean squared error (MSE) as the distillation loss which is weak for knowledge transfer. It can not effectively measure the relative entropy between two probability distributions. Besides, MSE is used as the loss function of both task-specific supervised term and distillation term. These two loss terms have different optimization targets, one is the ground truth heatmap, the other is the predicted heatmap generated by the teacher. The conflict between two loss terms will deviate the optimization into a sub-optimal situation.

To alleviate those limitations, we investigate an online pose distillation approach for *efficient* pose estimation. The proposed method has two vital aspects for efficiency. *One is that we simplify the distillation procedure to one stage. The other one is that the proposed method significantly improves the pose estimation accuracy comparing with the original network.* The whole framework is constructed with a Feature Aggregation Unit (FAU) and multiple auxiliary branches, where each branch is treated as a student. The student branch can be both the same or heterogeneous architectures, making up the OKDHP-Balance and the OKDHP-Unbalance architectures respectively and enabling the customization of different compression rates. The teacher is established by the weighted ensemble of the predictions of all students through the FAU. The FAU here captures the multi-scale information to obtain the higher-quality target heatmap.

Besides, to transfer the pose structural knowledge, the pixel-wise KL divergence loss is utilized to minimize the discrepancy between the target heatmaps and the predicted ones. In the final deployment, the target single-branch network is acquired by simply removing redundant auxiliary branches from the trained multi-branch network, which doesn't introduce any test-time cost increase.

The main contributions of this paper are listed below.

- To our best knowledge, we are the first to propose the online pose distillation approach, which distills the pose structure knowledge in one-stage manner.
- Both balanced and unbalanced versions of OKDHP are introduced, which can customize the target network with different compressing rates.
- Extensive experiments validate the effectiveness of our proposed method on two popular benchmark datasets: MPII and COCO.

2. Related Work

Human Pose Estimation Classical human pose estimation approaches mainly adopt the technique of pictorial structures [14, 2, 23, 45, 46] and graphical models [50, 49, 7, 10]. With the rapid development of deep convolutional neural networks [28, 51, 17], approaches based on CNNs became popular in recent years [54, 57, 12, 39, 52, 3]. DeepPose proposed by Toshev *et al.* [55] was the first attempt to regress the coordinates of body parts directly and shows superior performance than classical approaches. Tompson *et al*. [54] learned body structures by jointly optimize the convnets and graphical models. CPM [57] incorporate convolutional networks into the pose machine framework for the task of human pose estimation and directly performs pose matching on the heatmaps. Newell *et al.* [38] stacked several hourglass modules to iteratively refine the predictions. Intermediate supervision is also used to produce accurate intermediate heatmaps and prevent gradient vanishes. The hourglass module is highly related to conv-deconv architecture [35, 48]. Features in this module are first pooled down to a low resolution, then are upsampled and fused with high-resolution features. Chu *et al.* [12] try to incorporate hourglass networks with attention mechanisms to learn and infer contextual representations. Yang *et al.* [61] further improved its performance by using the pyramid residual model.

In addition to heavy networks for highly accurate pose estimation, highly efficient pose estimation networks have also been studied to meet the needs of real applications. Cao *et al.* [4] introduced a real-time estimation network with two branches where one branch generates the heatmap predictions, while the other one generates part affinity field, then a greedy algorithm is used to group the joints to the corresponding person. Kocabas *et al.* [27] proposed pose residual network that takes as input keypoints and person detections then perform keypoints assignment. MultiPoseNet achieves similar accuracy to Mask-RCNN [16] while being at least 4x faster. Based on OpenPose [4], [42] uses a MobileNet [21] as a backbone network and adopt lightweight refinement stage to reduce computational cost.

Knowledge Distillation Originally introduced by Hinton *et al.* [19], knowledge distillation transfers knowledge in the form of soft predictions from a large and computational expensive model to a single computational efficient model through a learning procedure. When training the target student model, this method makes full use of the extra supervisory signal provided by the soft output of the teacher model. In FitNet [47], the student was forced to mimic the

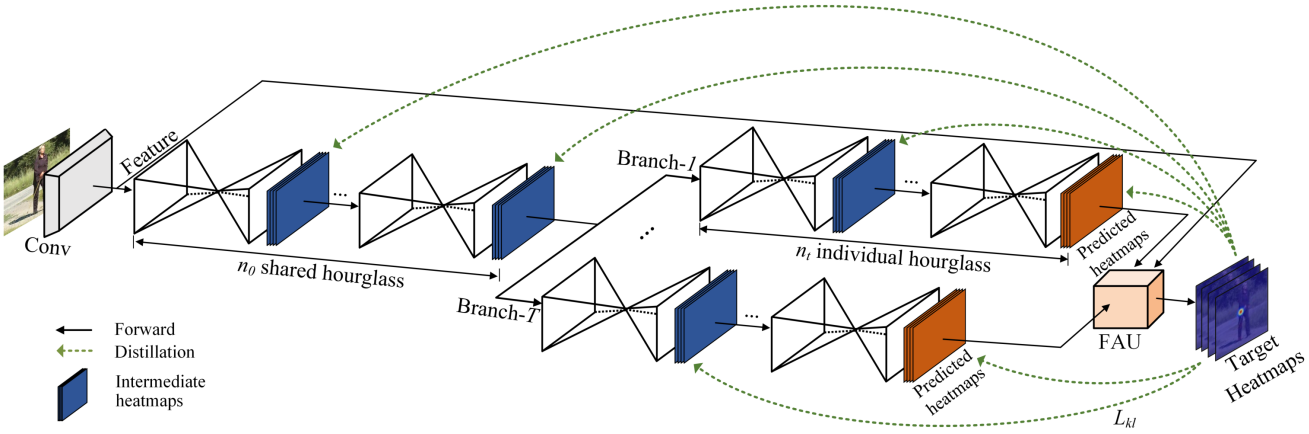

Figure 2. An overview of the proposed Online Knowledge Distillation for Human Pose estimation (OKDHP). Each branch serves as an independent pose estimator. The FAU learns to ensemble all branches to establish a stronger teacher model. L_{kl} denotes the KL divergence loss between intermediate heatmaps and ensemble heatmaps. We omit the conventional mean squared error loss L_{mse} for simplicity.

intermediate feature representations of the teacher. AT [64] try to transfer attention map of the teacher to the student. Kim *et al.* [25] introduces the paraphraser and translator network to assist the knowledge transfer procedure. In FSP [63], the student mimics the teacher's flow matrices, which are calculated as the inner product between feature maps from two layers. Traditional distillation methods always start with a powerful and cumbersome teacher model and perform one-way knowledge transfer to a compact student model. Online knowledge distillation [67, 68] simplifies the complex two-stage procedure by reducing the need for a pre-trained teacher model. ONE [68] builds a single multi-branch network and each branch learns from the ensemble results. Chen *et al.* [5] introduces the two-level distillation framework and uses a self-attention mechanism to construct diverse peer networks. Li *et al.* [32] made a further improvement to such branch-based network by enhancing the branch diversity.

Knowledge distillation methods have been widely used in many vision tasks, including object detection [30, 6, 13], line detection [20], semantic segmentation [62, 18, 34] and human pose estimation [66, 40, 56, 58]. DOPE [58] proposes to distill the 2D and 3D poses from three independent body part expert models to the single whole-body pose detection model. Nie *et al.* [40] distill the pose kernels via leveraging temporal cues from the previous frame in a one-shot feed-forward manner. Wang *et al.* distill the 3D pose knowledge from Non-Rigid Structure from Motion in weakly supervised learning. FPD [66] adopts the classical distillation approach and transfers the knowledge from an 8-Stack hourglass network to a lightweight 4-Stack hourglass network. Two shortcomings exist in the above work. A high-capacity teacher model is not always available and such complex two-stage learning will make the distillation inefficient.

3. Methodology

In this section, we first present a brief introduction to knowledge distillation, then we describe our proposed online knowledge distillation framework for efficient human pose estimation. Finally, we introduce the unbalanced version of our proposed OKDHP.

3.1. Teacher-Student Learning

Knowledge distillation [19], as one of the main model compression techniques [59, 37], follows the classic teacher-student learning paradigm. By treating a pre-trained heavy network as the teacher model, knowledge distillation aims to learn a lightweight student model, which is expected to master the expertise of the teacher, via transferring the knowledge from the teacher. Such distillation procedure can be formulated as:

$$L_{kd} = d(m_{stu}, m_{tea}), \tag{1}$$

where $d(\cdot)$ denotes the distance loss function, measuring the differences between two probability distributions. m_{stu} and m_{tea} represent the results generated by the student and the teacher, respectively.

With the task-specific supervised loss L_{task}, the whole loss function is given as:

$$L_{total} = L_{task} + \lambda L_{kd}, \tag{2}$$

where λ is the hyperparameter for balancing the two loss terms.

The vanilla knowledge distillation is a two-stage procedure where a cumbersome teacher model is first trained and fixed, and then the knowledge is distilled to a compact student model. This process increases the training complexity, making the distillation process inefficient.

3.2. Online Human Pose Distillation

To solve the problems in the vanilla distillation method, we propose an online knowledge distillation framework for efficient pose estimation. An overview of the proposed framework is illustrated in Fig. 2. The proposed OKDHP architecture contains a multi-branch network as the main network and an FAU module for building the teacher online. We adopt the Hourglass network [38] (HG) as our basic building block in the proposed framework, which is the most common block used in many state-of-the-art works [12, 24, 29].

3.2.1 The Main Network

The main network is in the multi-branch architecture that consists of T auxiliary homogeneous branches with the same network configuration (the same number of HG modules). That is, a total of T pose estimators are aggregated in the main network, each of which shares the first n_0 HG modules and is treated as a student. For every $1 \leq t \leq T$, branch-t has n_t individual HG modules. To make the method clearer, we firstly give the details of the OKDHP-balanced, where $n_1 = n_2 = ... = n_T$.

Thus, given a single RGB image, human pose estimation estimates a heatmap for each human anatomical keypoint, which represents the keypoint locations as Gaussian peaks. To train the multi-branch main network, we minimize the mean squared error (MSE) between the predicted heatmaps m_{pred} from each branch and the ground-truth heatmaps m_{gt}:

$$L_{mse} = \frac{1}{C} \sum_{t=0}^{T} \sum_{c=1}^{C} \|m_{gt}^{c} - m_{pred}^{c}(t)\|_2^2, \tag{3}$$

where C denotes the total number of human keypoints i.e. heatmap channels and T denotes the total number of network branches. $m_{pred}^{c}(t)$ is the predicted heatmap from branch-t at the c-th channel. Note that our network is built by stacking multiple hourglass modules, the supervision is applied not only on the final output but also on the intermediate heatmaps from each HG module.

3.2.2 The FAU Module

The Feature Aggregation Unit (FAU) learns to combine all the predicted heatmaps from T branches to establish a strong teacher model. The FAU module consists of multiple parallel transformations with different receptive fields, leveraging both local and global information to obtain accurate target heatmaps. The architecture of the proposed FAU is depicted in Fig. 3.

Previous image classification work [68] uses a simple conv block as the gate module to generate an importance score for each branch. But a simple conv block cannot effectively capture the contextual representation due to the body scale variation problem that exists in the natural scene. The multi-scale information is required to handle this problem. In this work, we focus on effectively capturing the multi-scale information to generate the target heatmaps with higher-quality. Inspired by the previous works [31, 15], we propose the FAU which is composed of multiple parallel transformations with different receptive fields. As shown in Fig. 3, multiple conv blocks in FAU have different receptive fields and are arranged in parallel. We take as input the features after the main network Conv block which contains more original information. The convolution operation starts with a small kernel size of 3, then consistently increases in the following branches (size of 3,5,7). In our network, additional 1×1 convolutions are mainly used as dimensional reduction methods to save computational resources. We further combine the average pooling of original inputs for richer representations. Then we concatenate all the splits and obtain the intermediate vector $\mathbf{v}$, denoted as:

$$\mathbf{v} = [v_{avg}, g([v_{conv3}, v_{conv5}, v_{conv7}])], \tag{4}$$

where $g(\cdot)$ denotes the global pooling function. v_{conv3}, v_{conv5}, v_{conv7} denote the results from the conv path with kernel size 3, 5 and 7, respectively. v_{avg} denotes the results from the average pooling path.

For any input feature maps, this configuration creates multi-scale features with each conv path that are aggregated to capture richer information for both local and global fields. We pass the intermediate vector $\mathbf{v}$ through the fully connected layer FC to fuse the information from different paths. Then, a channel-wise softmax operator is applied to obtain the soft attention vectors $a_{t,c}$. In the case of three network branches, for c-th heatmap we have

$$a_{1,c} + a_{2,c} + a_{3,c} = 1, \tag{5}$$

where c=1, 2, ..., C. Finally, we fuse predictions from multiple branches via an element-wise summation to obtain the weighted target heatmaps $\mathbf{m}_{tar}$:

$$\mathbf{m}_{tar} = \sum_{t=1}^{T} \mathbf{a}^t \otimes \mathbf{m}_s^t, \tag{6}$$

where $\mathbf{a}^t = [a_1, a_2, ..., a_c]^t$, $\mathbf{m}_s^t = [m_s^1, m_s^2, ..., m_s^c]^t$ and $\mathbf{m}_t \in \mathbb{R}^{H' \times W' \times C}$. Here, $\mathbf{a}^t$ is the weight for the t-th branch, $\mathbf{m}_s^t$ is the heatmaps generated by the t-th branch and $\otimes$ refers to the channel-wise multiplication between $\mathbf{a}^t$ and $\mathbf{m}_s^t$. Our experiments (see Section 4.4) prove that the weights generated by FAU can achieve better distillation performance.

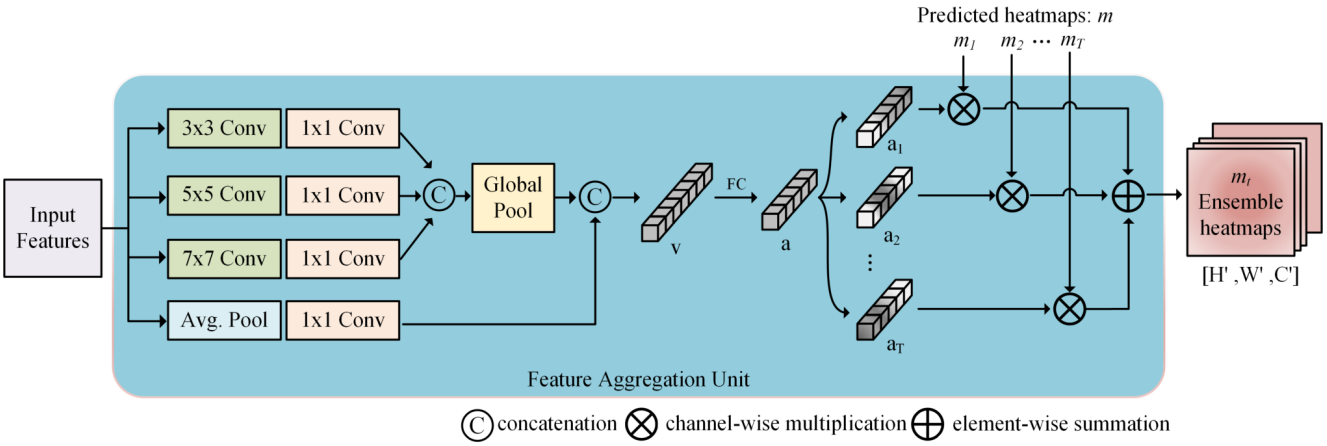

Figure 3. This module is proposed to effectively exploit the multi-scale information that can obtain target heatmaps with higher-quality. m_i denotes the predicted heatmaps that come from i-th branch. The final ensemble heatmaps are obtained by the weighted sum of heatmaps of all individual branches. Note that all conv blocks in the network are composed of regular convolutions, Batch Normalization and ReLU activation functions in sequence.

3.2.3 Pixel-wise Distillation

A proper distillation loss function is critical to the whole training procedure. Since pixel values on the heatmaps indicate the probabilities of pixels that belong to the keypoint. We align the heatmap generated by the student model with the target heatmaps. The target heatmaps obtained through the FAU play the role of a teacher model to teach each branch model (student) in our method. To transfer the pose structural knowledge, the pixel-wise Kullback-Leibler (KL) divergence loss is utilized to minimize the divergence between the heatmaps of the teacher model and the student model as follows:

$$L_{kl} = \frac{1}{W' \times H'} \sum_{i \in M} \sum_{t=0}^{T} KL(q^i_{tar}, q(t)^i_s), \qquad (7)$$

where W' and H' represent the heatmaps' width and height. q^i_{tar} and $q(t)^i_s$ denote the probabilities of the i-th pixel from the heatmap generated by the teacher model and the student model, respectively. $M = \{1, 2, ..., W' \times H'\}$ denotes all the pixels.

Overall To get a better understanding of our method, we describe the whole training procedure in Algorithm 1. For the proposed online human pose distillation method, the whole objective function consists of a conventional mean squared error loss L_{mse} for pose estimation and another loss term L_{kl} for online knowledge distillation:

$$L_{total} = \alpha L_{mse} + \beta L_{kl}. \qquad (8)$$

where α and γ are the hyperparameters to balance these two losses.

Algorithm 1 Online Human Pose Distillation

Input: Labelled Training dataset D; Training Epoch Number ϵ; Branch Number T; Network Structure $S \in \{S_{balance}, S_{unbalance}\}$

Output: Trained target pose estimate network θ^1 and other auxiliary estimators $\{\theta^i\}_{i=2}^T$

Initialize: Epoch e=1; Randomly initialize $\{\theta^i\}_{i=1}^T$

```
while e ≤ ϵ do
    Compute the heatmap predictions of all branches {θ^i}_{i=1}^T according to S;
    Compute the target heatmaps m_t with Eqn.(6) through FAU;
    Compute the MSE loss L_mse with Eqn.(3);
    Compute the distillation loss L_kl with Eqn.(7);
    Compute the total loss function with Eqn.(8);
    Update the model parameters {θ^i}_{i=1}^T;
    e=e+1;
end while
```

Model deployment: Use target pose estimator θ^1;

3.3. Unbalanced Architecture

To achieve a better distillation performance, a stronger teacher model is required. But in our balanced architecture, the teacher is fixed once we set up the target network. Here, the unbalanced variant is introduced to customize the student model with different compression rates, as shown in Fig. 1(c). For an unbalanced OKDHP architecture, each branch have different numbers of HG modules, where $n_1 \neq n_2 \neq ... \neq n_T$. For example, for a 3-branch unbalanced network, if a 4-Stack HG network is required for final deployment, the other two branches can

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Newell *et al.*(HG) [ECCV'16] [38]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	**90.9**
Ning *et al.* [TMM'17 [41]]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu *et al.* [CVPR'17] [12]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chen *et al.* [ICCV'17] [8]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.* [ICCV'17] [61]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al.* [ECCV'18] [24]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang *et al.* [ECCV'18] [53]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
FPD [CVPR'19] [66]	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
OKDHP	98.2	96.6	92.3	88.0	91.0	88.5	84.5	**91.7**

Table 1. Evaluation of our proposed OKDHP on MPII testing set (PCKh@0.5).

be set as a 6-stack and an 8-Stack network. Compared with the balanced structure, a better teaching performance can be achieved by utilizing the stronger representation ability of a larger network. Furthermore, the other benefit is that we can simultaneously obtain three different networks with comparable performances in one training procedure. This kind of network setting can be customized to other settings according to the actual needs. We demonstrate the effectiveness of unbalanced structure in Table 3 and present the detailed results in Table 5.

4. Experiments

To validate the effectiveness of our proposed method, we conduct several experiments on two popular human pose datasets, MPII [1] and COCO [33].

4.1. Implementation Details

Datasets The MPII dataset includes approximately 25K images containing over 40K subjects with annotated body joints, where 29K subjects are used for training and 11K subjects are used for testing. The images were collected using an established taxonomy of everyday human activities from YouTube videos. We adopted the same train/valid/test split as in [66]. Each person instance in MPII has 16 labeled joints.

The COCO keypoint dataset [33] presents naturally challenging imagery data with various poses. It contains more than 200k images and 250k person instances labeled with keypoints. In evaluation, we follow the commonly used train/val/test split. Each person instance is labeled with 17 joints.

Training details We implement all the methods in PyTorch [43]. For MPII, we resize the cropped images to 256×256 in pixels. Then we randomly augment the data with rotation degrees in [-30°, 30°], scaling with factors in [0.75, -1.25] and horizontal flip. For COCO, we resize the cropped image to 256×192 in pixels. Then we apply random horizontal flip, random rotation with degrees in [-40°, 40°] and random scale with factors in [0.7, 1.3]. We follow the standard data processing scheme for all images as in [66]. Adam [26] is used as the optimizer and we set the initial learning rate to 2.5e-4, weight decay to 1e-4. The learning rate is divided by 10 at 90 and 120 of the total 150 training epochs. We usually set α=1 and β=2 in Eqn.8. For the network architecture, we set the number of shared HG modules to half the number of total stacks. In the case of a 4-Stack OKDHP network, we have two shared HG modules and two individual HG modules for each branch. We set branch size to 3 as default. We use the OKDHP-balance architecture as our default scheme in the following experiments unless we specified. We adopt the official hourglass configurations as our baseline method in all experiments.

Evaluation Metric We use the standard Percentage of Correct Keypoint (PCK) metric which reports the percentage of correct keypoints lies within a normalized distance of ground truth. We use PCKh@0.5 for the MPII dataset, which refers to a threshold of 50% of the head diameter. For COCO, we use Object Keypoints Similarity (OKS) as our evaluation metric, which defines the similarity between different human poses.

OKDHP	Network	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	PCKh@0.5
×	2-Stack HG	96.7	95.3	89.2	84.0	87.8	83.9	79.5	88.6
✓		96.7	**95.4**	**89.9**	**84.1**	**89.0**	**84.7**	**81.1**	**89.2**
×	4-Stack HG	96.7	95.6	89.7	84.5	88.6	84.3	80.9	89.2
✓		**97.0**	**96.1**	**90.8**	**85.9**	**89.5**	**85.4**	**81.6**	**90.0**
×	8-Stack HG	96.9	95.9	90.6	86.0	89.8	86.0	82.5	90.2
✓		**97.3**	**96.1**	**91.2**	**86.8**	**89.9**	**86.9**	**83.1**	**90.6**

Table 2. PCKh@0.5 score of our proposed OKDHP on MPII validation set.

Method	PCKh@0.5	TrainCost
Baseline	89.2	14
FPD	89.7	66
OKDHP-Balance	90.0	47
OKDHP-Unbalance	90.2	64

Table 3. Comparison with different distillation methods based on the 4-Stack hourglass network on MPII validation set. TrainCost: Training cost in the unit of GFLOPS.

OKDHP	Network	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
×	2-Stack HG	71.7	90.5	78.4	69.0	75.8	74.6	91.9	80.6	71.6	79.2
✓		**72.8**	**91.5**	**79.5**	**69.9**	**77.1**	**75.6**	**92.5**	**81.5**	**72.5**	**80.3**
×	4-Stack HG	73.6	91.6	80.6	70.8	78.0	76.5	92.6	82.8	73.5	81.2
✓		**74.8**	**92.5**	**81.6**	**72.1**	**78.5**	**77.4**	**93.1**	**83.6**	**74.5**	**81.9**
×	8-Stack HG	75.3	91.6	82.6	73.0	79.1	78.0	92.9	84.0	75.2	82.3
✓		**76.2**	**92.6**	**83.7**	**73.4**	**80.2**	**78.8**	**93.6**	**85.2**	**75.9**	**83.3**

Table 4. Evaluation of our proposed OKDHP on COCO val2017 dataset.

OKDHP-Unbalance	Network	PCKh@0.5
Branch-1 (Target)	4-Stack HG	90.2
Branch-2 (Auxiliary)	6-Stack HG	90.3
Branch-3 (Auxiliary)	8-Stack HG	90.5

Table 5. Detailed results of the 3-branch OKDHP-Unbalance (two shared HG modules) network on MPII validation set.

4.2. Results on MPII dataset

We evaluate our method on the MPII dataset. Table 1 compares the PCKh@0.5 accuracy results of state-of-the-art methods and our proposed OKDHP on the MPII testing set. Table 2 reports the comparison of three varying-capacity networks trained by the conventional method and our proposed OKDHP. We can clearly observe that all networks benefit from our OKDHP training method, particularly for small networks achieving large performance gains. Specifically, our method improves various baseline networks ranging from 0.3 to 0.8. Considering that the performance of many state-of-the-art pose estimation networks, improves from 0.1% to 0.3% in PCKh scores, our performance is in fact, significant as compared to prior works. A 2-Stack HG network trained with OKDHP achieves similar performance with the original 4-Stack HG network but it only needs half the number of HG modules. In contrast to the conventional distillation method, a large pre-trained teacher is not necessary. We provide the visualized pose results in Fig. 4.

We compare our method with previous state-of-the-art distillation work FPD [66] and demonstrate the performance comparison of our proposed balanced and unbalanced structure in Table 3. The teacher network is an 8-Stack HG network with 90.2 PCKh@0.5 scores for FPD. We can clearly observe that both OKDHP balanced and unbalanced architectures outperforms the FPD method, validating the performance advantage of our method. The unbalanced structure outperforms the balanced structure by 0.2% but with 36%(17/47) FLOPS increase. OKDHP-Balance takes the least training cost, proves that our method is the most effective pose distillation approach.

We provide the detailed results of our proposed OKDHP-Unbalance architecture in Table 5. Three branches exists in our network. The first one is the target 4-Stack hourglass network for deployment. The other two branches plays the auxiliary roles, helping the target branch to achieve better performance.

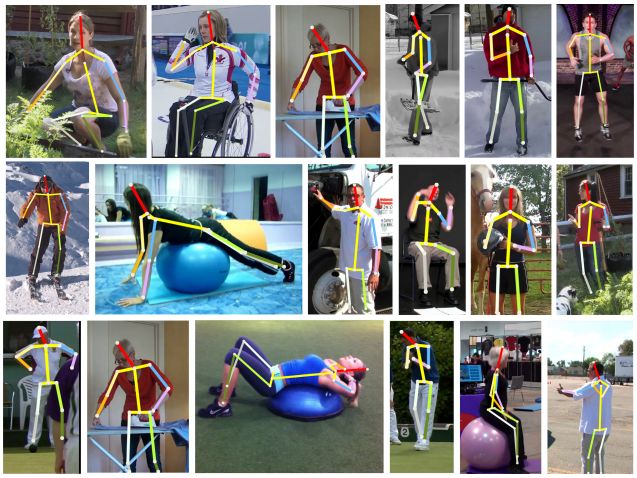

Figure 4. Visualized results on MPII dataset.

Loss	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	PCK
MSE	96.9	95.8	89.9	84.8	89.3	84.9	81.5	89.5
Ours	**97.0**	**96.1**	**90.8**	**85.9**	**89.5**	**85.4**	**81.6**	**90.0**

Table 6. Comparisons of different distillation loss functions on MPII validation set (PCKh@0.5).

4.3. Results on COCO dataset

Table 4 shows the results of the baseline method and our proposed OKDHP on the MS COCO keypoint dataset. In the test, a two-stage top-down paradigm is applied, same as in [60, 9, 65]. We use a detector with detection AP 56.4 for the person category on COCO val2017. From this table, we can observe that OKDHP method yields a more generalizable model compared to independent learning. This demonstrates that our method can still be applied to the large-scale dataset effectively.

4.4. Ablation Study

Loss Function The distillation loss function plays a critical role in the whole learning procedure. We compared the

performance of different distillation loss functions as shown in Table 6. FPD uses mean squared error (MSE) loss as the distillation term which is the same as the task-specific supervised term. We test the MSE loss in our proposed framework. In Table 6, this result shows that our proposed pixel-wise KL divergence is a better choice in comparison to the MSE. Our method can effectively distill the pose structural knowledge to enhance distillation performance.

Branch Size We evaluate the impact of branch size on the performance of our branch-based online pose distillation framework. Table 7 shows the performance on MPII validation set with varying branch sizes ranging from 2 to 5. We omit the case when branch size $n = 1$ since one branch cannot form the ensemble result. The baseline method denotes the vanilla 2-Stack networks without any modifications. We can clearly observe that OKDHP scales well with more branches and a 2-Stack OKDHP can be further improved if a larger branch size is allowed during training.

Branch Size	PCKh@0.5	#Params
Baseline	88.6	13.0M
2	89.2	15.5M
3	89.2	18.6M
4	89.3	21.7M
5	**89.4**	24.7M

Table 7. Impact of branch size for a 2-Stack OKDHP framework on MPII validation set (PCKh@0.5).

Individual HG Number We set $n_s = 2$ and $n_i = 2$ for a 4-Stack OKDHP network in the main experiment, indicating that we have two shared HG modules and two individual HG modules for each branch. Table 8 demonstrated the impact of the individual HG numbers on MPII validation set. From this table, we can see that if very few HG modules are shared, the performance will quickly drop. This brings the concept of branch diversity for such branch-based networks as mentioned in [5, 32]. Diversity will be hurt with the reducing number of individual HG modules, which will limit the effectiveness of within-group knowledge transfer. We usually set the number of shared and individual modules to half the number of total stacks to achieve the accuracy-efficiency trade-offs.

Individual HG numbers	1	2	3	4
PCKh@0.5	89.8	90.0	90.1	**90.2**
FLOPs	41G	47G	53G	59G

Table 8. Impact of the number of individual HG modules for a 4-Stack OKDHP network on MPII validation set (PCKh@0.5).

Sensitivity to Hyperparameter Table 9 demonstrates how the performance of our proposed framework is affected by the choice of hyperparameter β in Eqn.8. From this table, we can see that our method still has robust performance against varying β values ranging from 0.5 to 5.

β	0.5	1	2	3	4	5
PCKh@0.5	89.24	89.20	**89.28**	89.19	89.20	89.18

Table 9. Sensitivity to β for a 2-Stack OKDHP network on MPII validation set (PCKh@0.5).

FAU The goal of FAU is to generate accurate target heatmaps by weighted ensemble heatmaps from all auxiliary branches. To evaluate the effectiveness of our proposed FAU, we conduct various ablation studies on MPII validation set based on a 4-Stack HG network as shown in Table 10. We compare the performance of the following experiments. (1) Baseline: A vanilla 4-Stack HG network without any modification. (2) Mean: A simple average is applied to aggregate the heatmaps of all branches. (3) Gate: A simple attention module used in ONE [68] for the classification task. It was initially proposed for image classification. We reimplement this module so that it can be directly used in pose estimation network. (4) FAU: Our proposed module. We can see that FAU outperforms Mean and Gate by 0.3% and 0.2%, respectively. This confirms the usefulness of the FAU.

Baseline	Mean	Gate	FAU
89.2	89.7	89.8	**90.0**

Table 10. Ablation study on FAU module for a 4-Stack HG network on MPII validation set (PCKh@0.5).

5. Conclusion

In this paper, we propose a novel Online Knowledge Distillation framework by distilling Human Pose structure knowledge (OKDHP) in the one-stage manner. A network with multiple branches is utilized in the framework, where each branch is an independent pose estimator and is regarded as the student. The students from multiple branches are integrated into one teacher by the FAU module, which then optimizes the student branches in reverse. With OKDHP, the efficiency is significantly enhanced with reduced distillation complexity and improved model performance. Besides, the unbalanced OKDHP scheme is also introduced to enable the customization of the target network with different compression rates. Experiments have validated the effectiveness of our proposed OKDHP on two popular benchmark datasets.

Acknowledgement This work is supported by National Key Research and Development Project of China (Grant No.2017YFB1002803) and National Natural Science Foundation of China (Grant No.62072150).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2014.

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1014–1021. IEEE, 2009.

[3] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7035–7044, 2020.

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

[5] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3430–3437, 2020.

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

[7] Xianjie Chen and Alan Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. *arXiv preprint arXiv:1407.3399*, 2014.

[8] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1212–1221, 2017.

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

[10] Anoop Cherian, Julien Mairal, Karteek Alahari, and Cordelia Schmid. Mixing body-part sequences for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2360, 2014.

[11] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3218–3226, 2015.

[12] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

[13] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7032, 2019.

[14] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

[15] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[16] 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.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1013–1021, 2019.

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

[22] Dong-Hyun Hwang, Suntae Kim, Nicolas Monet, Hideki Koike, and Soonmin Bae. Lightweight 3d human pose estimation network training using teacher-student learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 479–488, 2020.

[23] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Citeseer, 2010.

[24] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 713–728, 2018.

[25] Jangho Kim, SeoungUK Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *arXiv preprint arXiv:1802.04977*, 2018.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[29] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11354–11361, 2020.

[30] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6356–6364, 2017.

[31] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019.

[32] Zheng Li, Ying Huang, Defang Chen, Tianren Luo, Ning Cai, and Zhigeng Pan. Online knowledge distillation via multi-branch diversity enhancement. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

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

[34] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[36] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[37] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

[38] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[39] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2108, 2018.

[40] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019.

[41] Guanghan Ning, Zhi Zhang, and Zhiquan He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2017.

[42] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv preprint arXiv:1811.12004*, 2018.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[44] Duc-Minh Pham. Human identification using neural network-based classification of periodic behaviors in virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 657–658. IEEE, 2018.

[45] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[46] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE international conference on Computer Vision*, pages 3487–3494, 2013.

[47] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[49] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3681, 2013.

[50] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *European Conference on Computer Vision*, pages 406–420. Springer, 2010.

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

[52] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[53] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 190–206, 2018.

[54] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a

graphical model for human pose estimation. *arXiv preprint arXiv:1406.2984*, 2014.

[55] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[56] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 743–752, 2019.

[57] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[58] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *European Conference on Computer Vision*, pages 380–397. Springer, 2020.

[59] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016.

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

[61] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *proceedings of the IEEE international conference on computer vision*, pages 1281–1290, 2017.

[62] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2824–2833, 2019.

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

[64] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

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

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

[67] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[68] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems*, pages 7517–7527, 2018.