Integral Knowledge Distillation for Multi-Person Pose Estimation

Xixia Xu, Qi Zou, Xue Lin, Yaping Huang, and Yi Tian

Abstract—**Both accuracy and efficiency are of equal importance to the human pose estimation. Most of the existing methods simply pursue excellent performance, sacrificing massive computing resources and memory. Out of this consideration, we present a novel compact and lightweight framework to train more efficient estimators using knowledge distillation. Three distillation mechanisms are proposed in our method from different perspectives, including logit distillation, feature distillation and structure distillation. Concretely, the logit distillation regards the output of teacher model as soft target to stimulate the student model. The feature distillation distills the high-level features of the teacher model to assist the student. Unlike the above strategies, the structure distillation considers the problem in a global view, aiming at ensuring the student prediction contains quite abundant structure knowledge like the teacher. We empirically demonstrate the effectiveness and efficiency of our methods on two multi-person pose estimation datasets (COCO and MPII). Specifically, our model can achieve competitive performance with the most state-of-the-art methods and consume only 35% model parameters and GFLOPs of our baseline (SimpleBaseline-ResNet-50) on the COCO dataset.**

Index Terms—**Multi-person pose estimation, knowledge distillation, feature distillation, structure distillation.**

I. Introduction

MULTI-PERSON pose estimation devotes to locate body parts of multiple persons in an image, such as keypoints on the arms, torsos, and the faces [1], [2]. The related tasks contain pose estimation in videos [3] and 3D pose estimation [4]. It has lots of potential related applications, such as human action recognition [5], human parsing [6]–[8].

In the recent years, convolutional neural network (CNN) [9] have achieved great success in human pose estimation. However, most existing methods achieved great performance sacrificing high computational cost and large memory. In computer vision field, various knowledge distillation strategies have been explored to train more lightweight and simple model without dramatic performance drop. Initially, knowledge distillation transfer soft target of the teacher model to a student model, which benefits the image classification [15]. Other strategies include distill the intermediate feature [16] of teacher model or transfer the attention information [17] from the teacher model to student. Inspired by this, the recent work [11](FPD) firstly introduced knowledge distillation to single-person pose estimation, but it doesn't work well for multi-person pose estimation. For multi-person pose estimation, the above normal distillation methods are weak at localizing fine-grained parts in images precisely, which directly degrade the accuracy of pose estimations. On the one hand, they overlook the importance of the high-level semantic information; on the other hand, they cannot better capture global structure information from the teacher model.

To overcome those above limitations, we formulate a novel Integral knowledge distillation (IKD) method for multi-person pose estimation. We take the high-level semantic information and high-order structural information into consideration during the knowledge transferring from teacher model to student one, which can estimate multi-person poses efficiently and effectively. Our model consists of three sub-modules, namely *logit distillation (LD)*, *feature distillation (FD)* and *structure distillation (SD)*. Firstly, inspired by the distillation methods applied in classification tasks, LD transfers output soft labels from the teacher model to the small one. Furthermore, given that the pose estimation is a structure prediction task, FD and SD are employed to further derive the abundant high-level semantic and structural knowledge from the teacher. Proverbially, high-level semantic features are crucial for keypoint prediction, while small student model is short on this. FD module is thus to ensure student model obtain high-level semantic information from teacher. Speaking of the SD, we deem that the teacher model's prediction contains quite reliable structure information. Therefore, we hope to transfer those global knowledge by adopting adversarial training, encouraging the predicted heatmaps from the teacher and student networks indistinguishable.

Our contributions are summarised in follows:

- We are the first to investigate the knowledge distillation strategies for training lightweight and compact estimation model for multi-person pose estimation.
- Based on the logit distillation, we explore two novel specialized distillation schemes: the feature distillation and structure distillation, enforcing the compact model to better capture high-level semantic information and global structure prediction from the large pose model.
- We conducted comprehensive evaluation to validate the efficacy and superiority of our method over most state-of-the-art approaches in the balance of model inference efficiency and performance on two commonly adopted benchmark datasets.

Manuscript received December 11, 2019; revised January 20, 2020; accepted February 16, 2020. Date of publication February 27, 2020; date of current version March 19, 2020. This work was supported by the Fundamental Research Funds for the Central Universities under Grant 2019JBM019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shiwen He. *(Corresponding author: Qi Zou.)*

The authors are with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: 19112036@bjtu.edu.cn; qzou@bjtu.edu.cn; 18112028@bjtu.edu.cn; yphuang@bjtu.edu.cn; tianyi@bjtu.edu.cn).

Digital Object Identifier 10.1109/LSP.2020.2975426

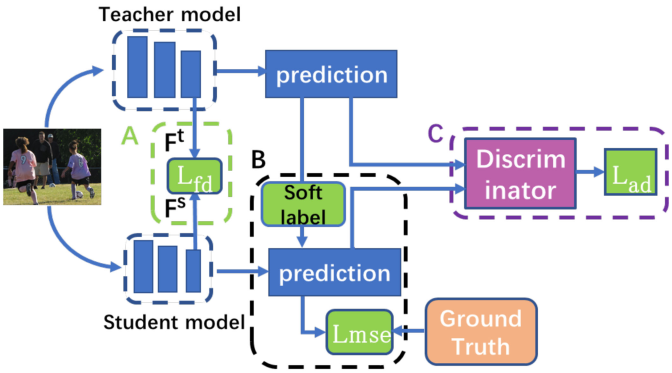

Fig. 1. Overview of the our integral knowledge distillation tailed for multi-person pose estimation. (A) Feature Distillation. (B) Logit Distillation. (C) Structure Distillation. In the training procedure, our teacher net is frozen and only the student net need to be trained, during which the discriminator net will be optimized. At test time, only our trained student model is used for pose estimation.

II. Proposed Approach

A. Framework Overview

In Fig. 1, we have the teacher and student model whose backbone networks adopt ResNet-50 and ResNet-18, respectively. The former has been trained with a heavy model and its parameter is frozen during the training. The latter owns smaller architecture and fewer parameters. The whole framework aims to produce a compact student network for pose estimation. To get such a network, an image is input into the res-block to extract features, during which FD translates the high-level semantic knowledge from the upper network to lower network. Then, the feature maps are sent into the decoder to obtain the heatmap prediction via upsampling. During this stage, LD transfers the teacher knowledge as a soft target to assist small network in objective function and SD captures global structure consistency between the prediction heatmaps from the teacher model and student model.

B. Integral Knowledge Distillation

Feature Distillation. The high-level feature representations possess significant semantic information that can provide more guidance for locating some challenging keypoints. The features coming from the last convolution layers of neural networks contain abundant semantic information, including discriminative and spatial information. These information are essential for localizing precise keypoints of human body. However, these information is hard to obtain for the small shallow network. Accordingly, we propose FD to transfer the above information from teacher network to student network, ensuring student model to learn discriminative and semantic features of images as the teacher model does.

To realize that, we measure the similarity of the feature maps that produced from the last residual block in teacher and student model respectively. Let f^t_{ij} denotes the feature similarity between the i-th pixel and the j-th pixel produced from the big model, and similarly f^s_{ij} denotes the similarity from the small model. We adopt the mean squared error to measure the matching of the high-level semantic knowledge between the teacher and student model. The formalized representation of the feature distillation loss is as below,

$$L_{fd} = \frac{1}{(W' \times H')^2} \sum_{i \in R} \sum_{j \in R} (f^s_{ij} - f^t_{ij})^2. \tag{1}$$

$R = \{1, 2, \ldots., W' \times H'\}$ denotes all the pixels of an image. The W' and H' represent the width and the height of the feature map. The similarity map between two pixels is just computed from the feature vectors F_i and F_j as

$$f_{ij} = \frac{F_i F_j}{(\| F_i \|_2)(\| F_j \|_2)}. \tag{2}$$

Logit Distillation. Our LD is to train the compact model from the soft targets or logits (predictions before keypoint localization) of the large model. The soft targets from teacher model carrys the abstract knowledge from the training dataset in advance, which can benefit student model in training. The define of L$_{soft}$ is as follows:

$$L_{soft} = \frac{1}{K} \sum_{K=1} \| h^s_k - h^t_k \|^2_2, \tag{3}$$

where the h^t_k and h^s_k represent the confidence maps for the k-th joint predicted by the pre-trained teacher model and the small target model, respectively. Here, the L_2 loss is chosen to make the L$_{soft}$ compatible with the original pose loss. The final loss of LD, which combines the original MSE loss using the ground truth labels with the L$_{soft}$ for matching the output of the teacher model inspired by [11] is defined as:

$$L_{ld} = \alpha L_{mse} + (1 - \alpha) L_{soft}, \tag{4}$$

where α is the balancing factor between the two loss terms. Its value is 0.5, estimated by cross-validation. As such, the small network learns from training samples by L_{mse} and simultaneously learn from the teacher model by L_{soft}.

Structure Distillation. Then, we focus on the global structure consistency of the teacher model and student model. The structure information here not only refers to the spatial contextual information but also the global distribution output. The student model isn't good at capturing these pivotal information, we therefore formulate to align this structure information between the teacher model and the student model.

Specifically, we adopt the adversarial learning to align the prediction heatmaps produced by the teacher and small model respectively. The compact model is regarded as a generator. The predicted heatmap H_s generated by it is regarded as the fake sample. We expect that H_s is approximate to H_t, which is the heatmap predicted by the teacher and is regarded as the real sample. The loss function of discriminator is defined as below to measure the matching of structural prediction,

$$L_{ad} = \mathbb{E}_{H_s \sim p_{s(H_s)}}[D(H_s)^2] + \mathbb{E}_{H_t \sim p_{t(H_t)}}[(D(H_t) - 1)^2] \tag{5}$$

where **1** is an all-one vector. $D(\cdot)$ tries to distinguish the H_t and H_s whist the small network tries to fool D in order to make H_s keep consistent with H_t. Eventually, D cannot distinguish them clearly and indicates that the small model has got the structure information from the teacher. D is a simple yet effective network with three convolution layers, which uses a 3×3 kernel with stride of 1, padding of 1, a BN layer and a ReLU activation function. We insert a non-local module into the final two layers to get the long-range spatial dependency.

Overall Optimization. The whole objective function for our pose knowledge distillation in the training phrase is:

$$L = \lambda L_{ld} + \beta L_{fd} + \gamma L_{ad}, \qquad (6)$$

where λ and β, γ are set as 0.5 and 0.2, 0.3 according to our experimental result. The optimization is conducted jointly through a minimax scheme that alternates between optimizing the compact network and the discriminator. We train the discriminator by minimizing L_{ad}. Given the well-trained D, the learning goal is to minimize the $L(G)$ relevant to the compact network:

$$L(G) = \lambda L_{ld} + \beta L_{fd} + \gamma L_{s}, \qquad (7)$$

where

$$L_s = \mathbb{E}_{H_s \sim p_{s(H_s)}} [(D(H_s) - 1)^2], \qquad (8)$$

is utilized to mitigate the difference between H_t and H_s.

III. Experiments

A. Experimental Setup

1) Datasets: We validate our model on two popular multi-person pose estimation datasets: COCO keypoint detection benchmark [13] and MPII Human Pose dataset [12]. For the COCO, its training set includes $57K$ images and $150K$ person instances. Ablation studies are validated on the COCO validate dataset (includes $5K$ images). The final comparison results are reported on the COCO test-dev dataset (includes $20K$ images). The MPII dataset consists of images taken from real-world activities with full-body pose annotations. There are about $25K$ images with $40K$ objects, where $12K$ objects are for testing and the remaining objects for training.

2) Performance Metrics: The public metric for the COCO is the standard OKS-based AP (average precision), where the OKS (object keypoints similarity) defines the similarity between the predicted heatmap and the groundtruth heatmap. For the MPII, the standard metric PCKh [12] (head-normalized probability of correct keypoint) score is utilized. The PCKh@0.5 score is reported in our result, 50% of the head size for normalization.

3) Implementation Details: We implement all the experiment in PyTorch. We select the ResNet-50 [9], 101, 152 as teacher models which are all initialized with the weights of the public Imagenet [14] pretrained model. For the COCO, the base learning rate is $5e{-}4$ train the teacher model for 130 epochs taking 84 hours with 2 TITAN GPU. For the MPII, we resize the images to 256 × 256 and train the teacher-50 (resnet) model for 120 epochs taking 9 hours. The training details of the whole framework are similar with the teacher model, except our ResNet-18 model is trained from scratch. During training, the parameters of the teacher model is frozen without updating. At test time, only the small and cost-effective target model is deployed without the heavy teacher network.

B. Comparisons to the State-of-the-Arts

1) Results on MPII Human Pose Dataset: Tab I shows the PCKh@0.5 results, the teacher and student model use ResNet-50 and ResNet-18 with the input size 256 × 256 respectively. From the table, our student net achieves a 87.6 PKCh@0.5 score without using any distillation strategy. Compared with it, our IKDNet surpasses it 2.8 points and achieves a comparable accuracy with the teacher model. It performs favorably against the hourglass model [18] and other methods [19]–[23]. Fig. 2 illustrates some qualitative results on the MPII dataset.

TABLE I
Results on MPII Test Set

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Insafutdinov *et al.* [24]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei *et al.* [25]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell *et al.* [18]	98.2	96.3	91.2	87.2	89.8	87.4	83.6	90.9
Sun *et al.* [21]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Ning *et al.* [10]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon *et al.* [26]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu *et al.* [19]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou *et al.* [27]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen *et al.* [28]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.* [23]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al.* [20]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang *et al.* [22]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
SBN *et al.* [29]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
TeacherNet	**98.5**	**96.6**	**91.9**	**87.6**	**91.1**	**88.1**	**84.1**	**91.5**
StudentNet	**96.7**	**94.6**	**87.9**	**82.9**	**87.6**	**82.1**	**75.8**	**87.6**
Ours	**98.1**	**96.1**	**91.0**	**87.1**	**90.1**	**87.4**	**83.6**	**90.8**

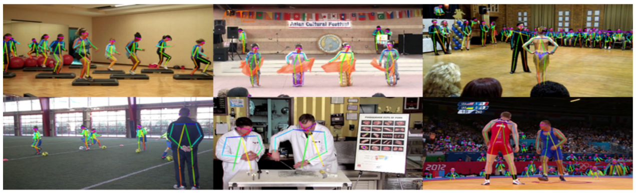

Fig. 2. Qualitative results of our model on the MPII dataset. Our model can achieve accurate result.

TABLE II
Comparison Results on the COCO Validation Set

Method	Params	GFLOPs	FPS	AP	AP^{50}	AP^{M}	AR
8-stage Hourglass [18]	25.1M	14.3	-	66.9	-	-	-
CPN [30]	27.0M	6.2	150	68.6	-	-	-
SBN [29]	34.0M	8.90	146	70.4	88.6	67.1	76.3
FPD [11]	3.0M	3.2	-	65.1	-	-	-
TeachNet-50	34.5M	8.9	143	70.5	89.4	68.1	77.3
TeachNet-101	53.0M	12.4	62	71.4	89.5	68.3	77.4
TeachNet-152	68.6M	15.7	21	72.0	89.7	68.6	77.6
Student-34	21.2M	5.3	161	67.5	85.3	64.5	72.5
Student-18	11.3M	2.8	182	65.1	82.2	61.4	70.2
Ours	**12.3M**	**3.2**	172	**70.3**	**87.1**	**66.8**	**75.1**

2) Results on COCO Keypoint Detection: We compare our IKDNet with existing popular methods on the COCO valid dataset (Table II) and test-dev dataset (Table III), respectively. In Table II, the input size is 256 × 192 and the backbone is ResNet-50 of all methods. In comparison to other methods, our IKDNet runs more efficiently and achieves the outstanding accuracy. And compared with the student model without any distillation, our method adds only a small fraction of parameters (about 10%) and gives a clear performance gain. Even with TeachNet-152, our model consumes only 30% parameters and FLOPs but achieves a competitive performance of less than 2 points decay, and a notable speed increase of about 8 times (172 FPS) faster than the TeachNet-152. It implies that our pose model is quite efficient than the top-performing architectures in theory and practice and can better meet the needs of practical applications. On the test-dev dataset, it also outperforms most of other counterparts despite with a smaller input size and still

TABLE III
COMPARISON OF RESULTS ON THE COCO TEST-DEV DATASET. **Top**: METHODS TRAINED ONLY WITH THE COCO TRAINVAL DATASET. **Middle**: RESULTS SUBMITTED TO THE COCO TEST-DEV LEADERBOARD [31]. * MEANS THAT THE METHOD INVOLVES EXTRA DATA FOR THE TRAINING. + INDICATES THE RESULTS USING THE ENSEMBLED MODELS. **Bottom**: THE RESULTS OF OUR SINGLE MODEL, TRAINED ONLY WITH THE COCO TRAINVAL DATASET. TEACHERNET REPRESENTS THE TEACHER MODEL OF OUR METHOD

Method	Backbone	Input Size	AP	AR
Associative Embedding [32]	-	512×512	65.5	70.2
G-RMI [33]	ResNet-101	353×257	64.9	69.7
CSANet [34]	ResNet-101	384×288	73.8	80.3
SBN [29]	ResNet-50	256×192	70.2	76.1
oks* [31]	-	-	72.0	77.1
bangbangren*+ [31]	ResNet-101	-	72.8	78.7
TeacherNet-50	ResNet-50	256×192	70.4	76.3
StudentNet-18	ResNet-18	256×192	64.5	70.1
Ours	**ResNet-18**	**256×192**	**70.3**	**74.1**

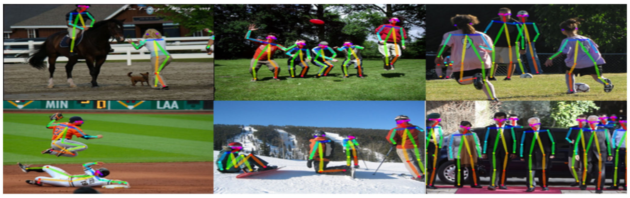

Fig. 3. Qualitative results of our model on the COCO dataset. Our model can precisely localize the keypoints inspite of extreme cases, complex and crowded poses.

TABLE IV
THE EFFECT OF DIFFERENT KNOWLEDGE DISTILLATION SCHEMES ON THE COCO 2017VAL DATASET

Method	**AP**	**FD**	**SD**	**LD**
S2: ResNet-18 [9]	65.1			
S2 + FD	66.7	√		
S2 + SD	67.4		√	
S2 + LD	66.1			√
Ours(ensembled)	70.3	√	√	√

generate satisfactory results compared with the teacher model. Notably, better results can be achieved with a larger input size like other methods in Table III. Fig. 3 illustrates some qualitative results on COCO dataset.

C. Ablation Study

In this section, we will describe the effectiveness of our proposed distillation schemes with different settings.

1) The Effectiveness of Our Distillations: The result in Table IV shows that the LD, FD and SD module brings in 1.0, 1.6 and 2.3 AP accuracy improvement respectively. For pose estimation, an improvement of 0.8 is a significant leap under the complex background [11]. It demonstrates that our distillation strategy is effective for the challenging poses at a cheaper cost. To explore how on earth this happens, we randomly select several samples with complex poses and visualize the results without any distillation (Fig. 4(3)) and with our method (Fig. 4(4)). With our SD, some highly challenging human postures are easier to handle (like bottom row in Fig. 4) since it‘s easier for the student model to align with the teacher's heatmap than directly learn from the groundtruth. Concretely, FD support the student model to acquire high-level semantic information from the teacher model thus it can better reduce the keypoint missing under clutter background (like upper row in Fig. 4). And LD and SD can make the student model learn more context and global structure information to localize human poses precisely on the whole.

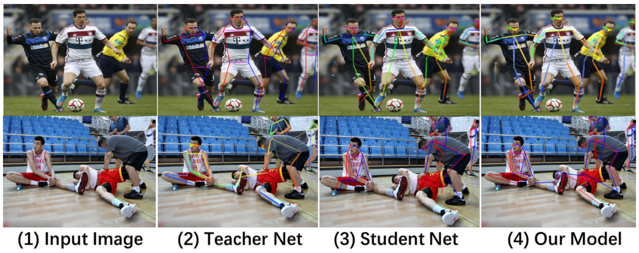

Fig. 4. Example results on our dataset. From left to right are: (1) Input images, (2) The results of the teacher net, (3) The results of the student net (4) The results of our method.

TABLE V
THE COMPARISON RESULTS ON COCO VAL2017. IMN: INITIALIZING THE NETWORK FROM THE WEIGHTS PRETRAINED ON IMAGENET DATASET

Backbone	**ResNet-18**	**ResNet-18 + ImN**
w/o distillation	64.2	65.1
+ LD	65.9	66.1
+ FD + LD	66.7	67.2
+ LD + AT	65.9	66.7
+ LD + FitNet	66.1	66.4
+ LD + SD	67.0	67.7
+ LD + SD + FD	**69.1**	**70.3**

2) Compare With Other Knowledge Distillations: To validate the superiority of our strategy, we make comparisons with other commonly adopted distillation schemes: KD [15], FitNet [16] and Attention Transfer (AT) [17]. KD is the basic distillation method equivalent to our LD. FitNet distills the intermediate features of the network. AT transfers the attention information. We mainly compare FD and SD with FitNet and AT respectively. Our FD and SD outperform FitNet and AT by 0.8 and 1.0 points respectively as in Table V, indicating our distillation strategies can better learn instructive information for keypoint localization compared with other common distillations. IKDNet with integrated three components achieves the best localization performance. The underlying reasons may be that keeping structure consistency with the teacher model through adversarial learning, and emphasizing high-level semantic information are specially suitable for such tasks as fine-grained keypoint localization.

IV. CONCLUSION

In this letter, we present a novel compact framework to train more efficient and effective pose estimator with exclusive knowledge distillation schemes nearly without accuracy decrease. Basically, we use a general soft label of the large network to assist in small network training. In the feature level, we transfer the high-level features to support the student network. Considering the structure consistency, we employ the adversarial learning to align the global predictions between the large and small model. Extensive experiments verify the effectiveness of our idea with state-of-the-art results on two popular multi-person pose estimation datasets. And we also hope our methods could inspire more ideas on the lightweight multi-person pose estimation field.

REFERENCES

[1] E. Cho and D. W. Kim, “Accurate human pose estimation by aggregating multiple pose hypotheses using modified kernel density approximation,” *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 445–449, 2015.

[2] J. Yan, S. Shen, Y. Li, and Y. Liu, “An optimization based framework for human pose estimation,” *IEEE Signal Process. Lett.*, vol. 17, no. 8, pp. 766–769, Aug. 2010.

[3] F. Zhou and F. D. La Torre, “Spatio-temporal matching for human pose estimation in video,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1492–1504, Aug. 2016.

[4] B. Babagholamimohamadabadi, A. Jourabloo, A. Zarghami, and S. Kasaei, “A bayesian framework for sparse representation-based 3-D human pose estimation,” *Signal Process. Lett.*, vol. 21, no. 3, pp. 297–300, 2014.

[5] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, “Skeletonnet: Mining deep part features for 3-D action recognition,” *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, Jun. 2017.

[6] X. Liang and K. Gong. X. Shen, and L. Lin, “Look into person: Joint body parsing & pose estimation network and a new benchmark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019.

[7] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan, “Towards unified human parsing and pose estimation,” in *Proc. Comput. Vision Pattern Recognit.*, 2014, pp. 843–850.

[8] X. Nie, J. Feng, and S. Yan, “Mutual learning to adapt for joint human parsing and pose estimation,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 519–534.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[10] G. Ning, Z. Zhang, and Z. He, “Knowledge-guided deep fractal neural networks for human pose estimation,” *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.

[11] F. Zhang, X. Zhu, and M. Ye, “Fast human pose estimation,” in *Proc. Comput. Vision Pattern Recognit.*, 2018, pp. 3517–3526.

[12] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, “2-D human pose estimation: New benchmark and state of the art analysis,” in *Proc. Comput. Vision Pattern Recognit.*, 2014, pp. 3686–3693.

[13] T. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[14] O. Russakovsky *et al.*, “Imagenet large scale visual recognition challenge,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 211–252.

[15] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. NIPS Deep Learn. Workshop*, 2014.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[17] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” in *Proc. Int. Conf. Learn. Represent.*, 2017.

[18] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Proc. Eur. Conf. Comput. Vision.*, 2016, pp. 483–499.

[19] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, “Multi-context attention for human pose estimation,” in *Proc. Comput. Vision Pattern Recognit.*, 2017, pp. 5669–5678.

[20] L. Ke, M. Chang, H. Qi, and S. Lyu, “Multi-scale structure-aware network for human pose estimation,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 731–746.

[21] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang, “Human pose estimation using global and local normalization,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5600–5608.

[22] W. Tang, P. Yu, and Y. Wu, “Deeply learned compositional models for human pose estimation,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 197–214.

[23] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning feature pyramids for human pose estimation,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1290–1299.

[24] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “Deepercut: A deeper, stronger, and faster multi-person pose estimation model,” in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 34–50.

[25] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in *Proc. Comput. Vision Pattern Recognit.*, 2016, pp. 4724–4732.

[26] D. C. Luvizon, H. Tabia, and D. Picard, “Human pose regression by combining indirect part detection and contextual information,” *Comput. Graph.*, vol. 85, pp. 15–22, 2019.

[27] C. Chou, J. Chien, and H. Chen, “Self adversarial training for human pose estimation,” in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 17–30.

[28] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang, “Adversarial posenet: A structure-aware convolutional network for human pose estimation,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1221–1230.

[29] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 472–487.

[30] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in *Proc. Comput. Vision Pattern Recognit.*, 2018, pp. 7103–7112.

[31] MS-COCO, “Coco keypoint leaderboard,” 2014. [online]. Available: http://cocodataset.org/

[32] A. Newell, Z. Huang, and J. Deng, “Associative embedding: End-to-end learning for joint detection and grouping,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 2277–2287.

[33] G. Papandreou *et al.*, “Towards accurate multi-person pose estimation in the wild,” in *Proc. Comput. Vision Pattern Recognit.*, 2017, pp. 3711–3719.

[34] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, “Multi-person pose estimation with enhanced channel-wise and spatial information,” in *Proc. Comput. Vision Pattern Recognit.*, 2019, pp. 5667–5675.

[35] F. Xia, P. Wang, X. Chen, X. Geng, and A. Yuille, “Joint multi-person pose estimation and semantic part segmentation,” in *Proc. Comput. Vision Pattern Recognit.*, 2017, pp. 6080–6089.

[36] G. Papandreou, T. Zhu, L. Chen, J. Tompson, and K. Murphy, “PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 282–299.