Effective Whole-body Pose Estimation with Two-stages Distillation

Zhendong Yang[1,2*] Ailing Zeng[2] Chun Yuan[1†] Yu Li[2†]
[1]Tsinghua Shenzhen International Graduate School
[2]International Digital Economy Academy (IDEA)
yangzd21@mails.tsinghua.edu.cn
yuanc@sz.tsinghua.edu.cn {zengailing, liyu}@idea.edu.cn

Abstract

*Whole-body pose estimation localizes the human body, hand, face, and foot keypoints in an image. This task is challenging due to multi-scale body parts, fine-grained localization for low-resolution regions, and data scarcity. Meanwhile, applying a highly efficient and accurate pose estimator to widely human-centric understanding and generation tasks is urgent. In this work, we present a two-stage pose **D**istillation for **W**hole-body **P**ose estimators, named **DWPose**, to improve their effectiveness and efficiency. The first-stage distillation designs a weight-decay strategy while utilizing a teacher's intermediate feature and final logits with both visible and invisible keypoints to supervise the student from scratch. The second stage distills the student model itself to further improve performance. Different from the previous self-knowledge distillation, this stage finetunes the student's head with only 20% training time as a plug-and-play training strategy. For data limitations, we explore the UBody dataset that contains diverse facial expressions and hand gestures for real-life applications. Comprehensive experiments show the superiority of our proposed simple yet effective methods. We achieve new state-of-the-art performance on COCO-WholeBody, significantly boosting the whole-body AP of RTMPose-l from 64.8% to 66.5%, even surpassing RTMPose-x teacher with 65.3% AP. We release a series of models with different sizes, from tiny to large, for satisfying various downstream tasks. Our code and models are available at* https://github.com/IDEA-Research/DWPose.

1. Introduction

Whole-body pose estimation plays a crucial role in numerous human-centric perception, understanding, and generation tasks, including 3D whole-body mesh recovery [1, 25, 31, 65], human-object interaction [7, 42], and pose-conditioned human image and motion generation [10, 24, 27, 59]. Furthermore, capturing human poses for virtual content creation and VR/AR has gained significant popularity, relying on user-friendly algorithms like OpenPose [2] and MediaPipe [29, 62]. Despite the convenience of these tools, their performance remains unsatisfactory, limiting their potential. Therefore, further advancements in human pose estimation technology are essential to fully unleash the potential of user-driven content creation. Compared with human pose estimation with body-only keypoints detection, whole-body pose estimation faces more challenges from 1) the hierarchical structures of the human body for fine-grained keypoints localization; 2) the small resolutions of hand and face; 3) the complex body parts matching for multiple persons in an image, especially for occlusion and complex hand poses; 4) data limitation, especially for diverse hand pose and head pose for the whole-body images.

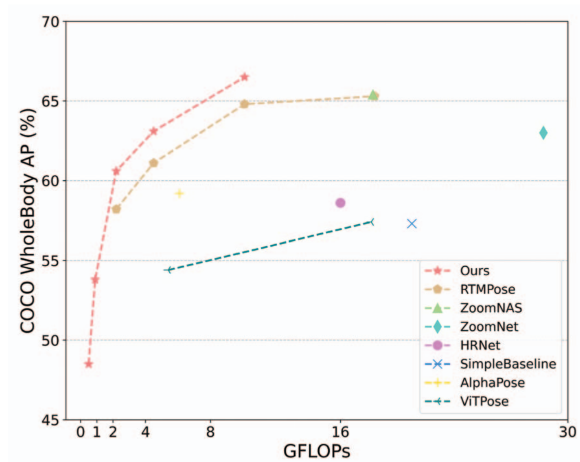

Figure 1. Comparison of our model and related models for whole-body pose estimation on COCO-WholeBody.

Besides, before deploying a model, it is essential to compress it into a lightweight network. The basic compression tools comprise distillation [14], pruning [8], and quanti-

*This work was done when Zhendong was an intern at IDEA.
†Corresponding authors.

zation [60]. Knowledge distillation (KD) is proposed to enhance the efficiency of a compact model without incurring extra costs during inference. This technique enables a student to inherit knowledge from a larger teacher and has found widespread application in various tasks, such as classification [69], detection [49], and segmentation [28].

In this paper, we explore KD for whole-body pose estimation to benefit many downstream applications, resulting in a series of real-time pose estimators with high performance and efficiency. Specifically, we propose a novel two-stages pose distillation framework DWPose, which achieves state-of-the-art performance, as shown in Fig. 1. We adopt the latest pose estimator RTMPose [18] as the basic model, which has been trained on COCO-WholeBody [19, 26].

In the first-stage distillation, we natively leverage the teacher's (*e.g.*, RTMPose-x) intermediate layer and final logits to guide the student model (*e.g.*, RTMPose-l). Previous pose training distinguishes keypoints via visibility and only uses visible keypoints for supervision. Unlike that, we use the teacher's complete outputs with both visible and invisible keypoints as final logits, which can impart reasonable and comprehensive values to facilitate the student's learning process. Meanwhile, we employ a weight-decay strategy to enhance the efficacy, gradually reducing the distillation's weight throughout the entire training phase. Due to a better head will determine a more precise localization, the second-stage distillation proposes a head-aware self-KD to enhance the capacity of the head. We construct two identical models and select one as the teacher and the other as the student to be updated. The student backbone is frozen, and only its head is updated through the logit-based distillation. Notably, this plug-and-play approach allows the student to achieve better results with 20% training time, whether trained from scratch with distillation or without, and can be used for any dense prediction heads.

Data volume and diversity addressing different scales of human body parts will affect the model performance. Suffering from the limited holistic annotated keypoints on existing datasets, existing estimators fail to localize well on fine-grained fingers and face landmarks. Thus, we explore the data impact by incorporating an additional UBody [25] dataset, primarily comprising diverse face and hand keypoints captured in various real-life scenes.

Therefore, our contributions can be summarized as:

- We introduce a two-stage pose knowledge distillation method, pursuing efficient and precise whole-body pose estimation.
- To break the whole-body data limitation, we explore more comprehensive training data, especially on diverse and expressive hand gestures and facial expressions, making it practical for real-life applications.
- Based on the latest RTMPose as our base model, our proposed distillation and data strategies can significantly improve RTMPose-l from 64.8% to 66.5% AP, even surpassing RTMPose-x teacher with 65.3% AP. We also validate the powerful effectiveness and efficiency of DWPose on the generation task.

2. Related work

2.1. 2D Whole-body Pose Estimation

This task targets locating expressive body, hand, feet, and face keypoints for all persons in an image simultaneously [2, 13, 19]. Due to the lack of whole-body annotations, most previous models are designed for body-only [22, 40, 47, 53], hand-only [6, 32, 62, 68], or face-only [21, 46, 67]. Openpose [3] combines different datasets for separate body parts. MediaPipe [29, 62] builds a perception pipeline for easy-to-use applications, especially for whole-body landmark detection. With the emergence of whole-body data [9, 19], the models for whole-body pose estimation make great progress [13, 18, 48]. Specifically, ZoomNet [19] proposes the first top-down method with a hierarchical single network to solve the scale variance of different body parts. ZoomNAS [48] further explores a neural architecture search framework for jointly searching the model architecture and the connections between different sub-modules to promote both accuracy and efficiency. TCFormer [61] introduces progressive clustering and merging vision tokens for various locations, sizes, and shapes in multiple stages, preserving different scale information well. Recently, RTMPose [18] has discussed key factors in pose estimation and built a real-time model, achieving state-of-the-art results on COCO-WholeBody. However, it still suffers from redundant model designs and data limitations, especially for diverse hand and face poses.

2.2. Knowledge Distillation

Knowledge distillation is a way to compress the model. Hinton *et al*. [14] first proposed to supervise the student with the soft labels from the teacher's output. The method is originally designed for classification and is also called logit-based distillation. Some following works [15, 52, 54] utilize teacher's logits in different ways, transferring more knowledge from soft labels, target and non-target logits [16, 58, 64, 66, 58]. From the logit-based distillation to feature-based distillation, the knowledge is transferred from intermediate layers [17, 55, 57] and it extends the distillation to various tasks, including detection [4, 56], segmentation [38], generation [30] and so on.

Utilizing KD in human pose estimation has been rarely studied [23, 35, 45, 50]. Existing works either distill the heavy heatmaps for body-only pose estimation [23, 35] or focus on gathering separate body-part experts' knowledge

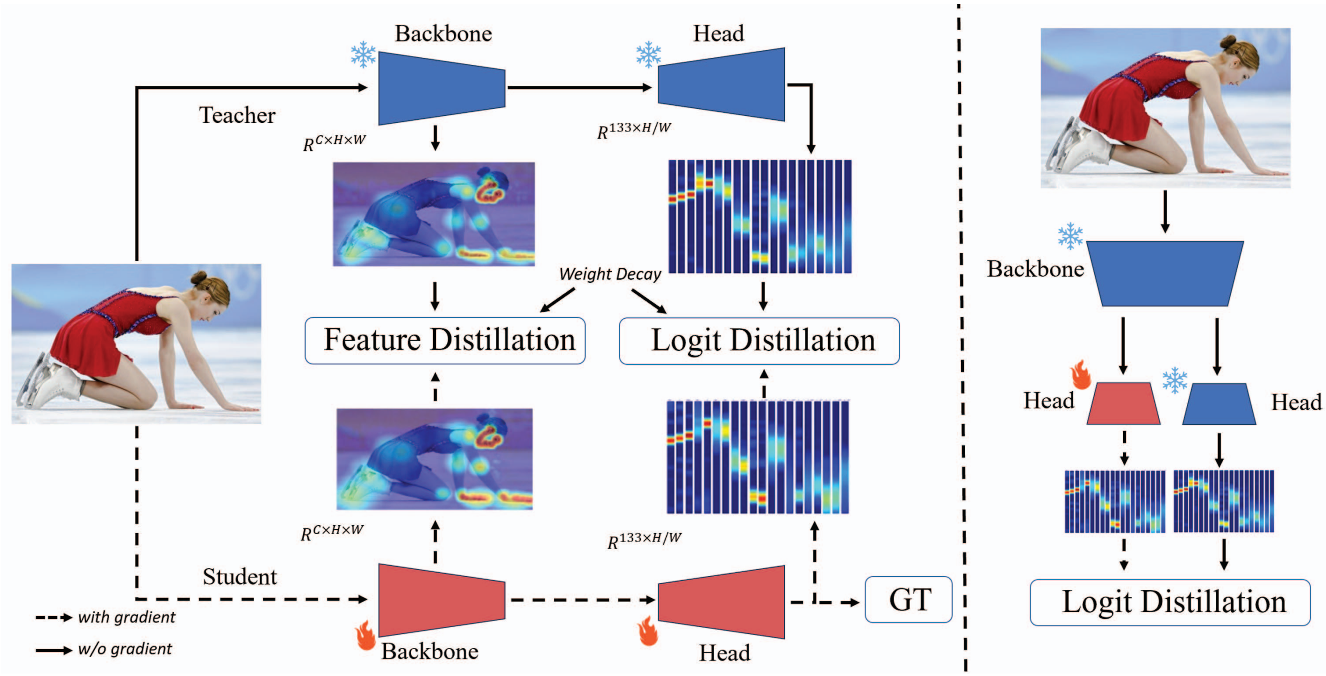

Figure 2. Pipeline of Two-stages Pose Distillation (TPD). On the left, the first-stage distillation adopts a traditional style leverage both feature and logit level. On the right, the second-stage distillation employs the student itself to teach a new head for enhanced performance.

into a single deep network designed for whole-body 2D-3D pose detection [45]. 2D whole-body pose estimation is a basic task for 3D pose estimation and is more holistic than body-only pose estimation. Our proposed DWPose is the first work to explore efficient KD strategies for this task.

3. Method

In the following, we provide a detailed exposition of the two-stage pose distillation (TPD). As shown in Fig. 2, it comprises two distinct stages. The first-stage distillation involves a pre-trained teacher guiding the student from scratch at both the feature and logit levels. On the other hand, the second-stage distillation can be considered a self-KD approach. The model employs its own logits to train its head without any labeled data, leading to significant performance enhancements within a concise training period.

3.1. The First-stage distillation

We denote the feature from the teacher's and student's backbone as F^t and F^s, and the teacher and student's final output logit as T_i and S_i. The first-stage distillation forces the student to learn the teacher's feature F^t and logit T_i.

3.1.1 Feature-based distillation

For the feature-based distillation, we force the student to mimic the teacher's layer from the backbone directly. We utilize MSE loss to calculate the distance between the student's feature F^s and the teacher's feature F^t. To learn the knowledge from the teacher's feature map, the distillation loss of the feature can be formulated as:

$$L_{fea} = \frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(F^t_{c,h,w} - f(F^s_{c,h,w})\right)^2, \quad (1)$$

where f is a 1×1 convolutional layer to reshape the F^s to the same dimension as F^t. H, W, C denote the height, width and channel of the teacher's feature.

3.1.2 Logit-based distillation

RTMPose [18] predicts pose keypoints with a SimCC-based [22] algorithm that treats keypoint localization as a classification task for horizontal and vertical coordinates. Following this design, we can also apply the logit-based knowledge method to it. To begin with, we review the original classification loss for RTMPose as follows:

$$L_{ori} = -\sum_{n=1}^{N}\sum_{k=1}^{K} W_{n,k}\cdot\sum_{i=1}^{L}\frac{1}{L}\cdot V_i log(S_i), \quad (2)$$

where N is the number of the person samples in a batch, K is the number of keypoints, *e.g.*, 133 for COCO-WholeBody [19], L is the length of the x or y localization bins. $W_{n,k}$ is a target weight mask to distinguish invisible keypoints. V_i is the label value.

For the logit-based distillation, we follow the form of the original loss L_{ori}. It's worth noting that we drop the target weight mask W for distillation. Different from the label value, the invisible keypoints can also be distributed a reasonable value by the teacher. So we argue such value is also helpful, and we also verify it in Sec. 5.6. The distillation loss of the logits can be formulated as:

$$L_{logit} = -\frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{i=1}^{L} T_i log(S_i). \tag{3}$$

3.1.3 Weight-decay strategy for distillation

With feature distillation loss L_{fea} and logits distillation loss L_{logit}, we can train the student with the total loss as:

$$L = L_{ori} + \alpha L_{fea} + \beta L_{logit}, \tag{4}$$

where α and β are hyper-parameters to balance the loss.

Inspired by a detection distillation method TADF [41], we apply a weight-decay strategy for the distillation to reduce the distillation penalty gradually. This strategy helps the student to focus more on the label and achieve better performance. We utilize a time function $r(t)$ to implement the strategy, which is as follows:

$$r(t) = 1 - (t-1)/t_{max}, \tag{5}$$

where $t \in (1, ..., t_{max})$ is the current epoch and t_{max} is the total epochs for training. Then the final loss for the first-stage distillation can be formulated as:

$$L_{s1} = L_{ori} + r(t) \cdot \alpha L_{fea} + r(t) \cdot \beta L_{logit}, \tag{6}$$

3.2. The Second-stage distillation

In the second distillation stage, we try to utilize the trained student model to teach itself for a better performance. In this way, it can bring improvements for the students, whether trained from scratch with distillation or not.

The pose estimator comprises the encoder (backbone) and decoder (head). Based on the trained model, we first build a student with a trained backbone and an untrained head. The teacher is the same model with a trained backbone and head. During training, we freeze the student's backbone and update the head. Because the teacher and the student have the same architecture, we only need to extract the feature from the backbone once. Then, the feature is fed into the teacher's trained head and the student's untrained head to get the logits T_i and S_i, respectively. Following the form in Eq. 3, we train the student with L_{logit} for the second-stage distillation. It's worth noting that we drop the original loss L_{ori}, which is calculated with label value. Using γ to denote the hyper-parameter for loss scale, the final loss for the second-stage distillation can be formulated as:

$$L_{s2} = \gamma L_{logit}. \tag{7}$$

Different from previous self-KD methods, our proposed head-aware distillation can efficiently distill the knowledge from the head with only 20% training time and further improve the localization capability.

4. Experiments

4.1. Datasets and Details

Datasets. We conduct experiments on COCO [19, 26] and UBody [25]. For the COCO dataset, we follow the standard splitting of train2017 and val2017, which use the 118K train images for training and 5K val images for testing, respectively. Unless specifically, we adopt a commonly used person detector provided by SimpleBaseline [47] with 56.4% AP for the COCO val dataset. UBody consists of over 1M frames from 15 real-life scenarios. It provides the corresponding 133 2d keypoints and SMPL-X parameters. Notably, the original dataset only focuses on 3D whole-body estimation and does not validate the effectiveness of 2D annotations. we pick every frame at an interval of 10 frames from the video used for both training and testing.

Implementation details. For the first-stage distillation, we utilize two hyper-parameters α and β in Eq. 6 to balance the loss scale. For all the experiments, we adopt $\{\alpha = 0.00005, \beta = 0.1\}$ on both COCO and UBody. The second-stage distillation has one hyper-parameter γ to balance the loss scale in Eq. 7. For all the experiments, we adopt $\gamma = 1$. The training setting, such as the optimizer, learning rate, and training epochs for the first-stage distillation, is the same as training the student without distillation [18]. For the two-stage distillation, we only need a short training time of about 1/5 of the whole training epochs. The other training settings still remain the same. This early stopping method helps to save much time for training. We use 8 GPUs to conduct the experiments with MMPose [5] based on Pytorch [36]. As a top-down pose estimator following RTMPose, we use the person detection boxes with 56.4 AP on the COCO val2017 dataset and the provided ground-truth box on Ubody.

4.2. Main Results

For a fair comparison, we evaluate our models on the public COCO-WholeBody dataset. As shown in Tab. 1, we utilize the larger RTMPose-x and RTMPose-l as the teacher to guide DWPose-l and the other student models, respectively. With our TPD, the models with different sizes and input resolutions all achieve significant improvements. Specifically, DWPose-m gets 60.6 whole AP with 2.2 GFLOPs. The performance is 4.1% higher than the baseline, while the consumption for the inference still remains the same, making it friendly to deploy. Interestingly, DWPose-l achieves 63.1 and 66.5 whole AP under two different input resolutions, which both beat the teacher

	Method	Input Size	GFLOPs	whole-body		body		foot		face		hand	
				AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Whole-body	SN† [13]	N/A	N/A	32.7	45.6	42.7	58.3	9.9	36.9	64.9	69.7	40.8	58.0
	OpenPose [2]	N/A	N/A	44.2	52.3	56.3	61.2	53.2	64.5	76.5	84.0	38.6	43.3
Bottom-up	PAF† [3]	512×512	329.1	29.5	40.5	38.1	52.6	5.3	27.8	65.6	70.1	35.9	52.8
	AE [34]	512×512	212.4	44.0	54.5	58.0	66.1	57.7	72.5	58.8	65.4	48.1	57.4
Top-down	DeepPose [43]	384×288	17.3	33.5	48.4	44.4	56.8	36.8	53.7	49.3	66.3	23.5	41.0
	SimpleBaseline [47]	384×288	20.4	57.3	67.1	66.6	74.7	63.5	76.3	73.2	81.2	53.7	64.7
	HRNet [40]	384×288	16.0	58.6	67.4	70.1	77.3	58.6	69.2	72.7	78.3	51.6	60.4
	PVT [44]	384×288	19.7	58.9	68.9	67.3	76.1	66.0	79.4	74.5	82.2	54.5	65.4
	FastPose50-dcn-si [9]	256×192	6.1	59.2	66.5	70.6	75.6	70.2	77.5	77.5	82.5	45.7	53.9
	ZoomNet [19]	384×288	28.5	63.0	74.2	74.5	81.0	60.9	70.8	88.0	92.4	57.9	73.4
	ZoomNAS [48]	384×288	18.0	65.4	74.4	74.0	80.7	61.7	71.8	88.9	93.0	62.5	74.0
	ViTPose+-S [51]	256×192	5.4	54.4	-	71.6	-	72.1	-	55.9	-	45.3	-
	ViTPose+-H [51]	256×192	122.9	61.2	-	75.9	-	77.9	-	63.3	-	54.7	-
	RTMPose-m	256×192	2.2	58.2	67.4	67.3	75.0	61.5	75.2	81.3	87.1	47.5	58.9
	RTMPose-l	256×192	4.5	61.1	70.0	69.5	76.9	65.8	78.5	83.3	88.7	51.9	62.8
	RTMPose-l	384×288	10.1	64.8	73.0	71.2	78.1	69.3	81.1	88.2	91.9	57.9	67.7
	RTMPose-x	384×288	18.1	65.3	73.3	71.4	78.4	69.2	81.0	88.8	92.2	59.0	68.5
	RTMPose-l + UBody	256×192	4.5	62.1	70.6	69.7	76.9	65.5	78.1	84.1	89.3	55.1	65.4
	RTMPose-l + UBody	384×288	10.1	65.4	73.2	71.0	77.9	68.6	80.2	88.5	92.2	60.6	69.9
	DWPose-t	256×192	0.5	48.5	58.4	58.5	67.0	46.5	63.6	73.5	80.7	35.7	49.0
	DWPose-s	256×192	0.9	53.8	63.2	63.3	71.3	53.3	69.0	77.6	84.1	42.7	54.9
	DWPose-m	256×192	2.2	60.6	69.5	68.5	76.1	63.6	77.2	82.8	88.1	52.7	63.4
	DWPose-l	256×192	4.5	63.1	71.7	70.4	77.7	66.2	79.0	84.3	89.4	56.6	66.5
	DWPose-l	384×288	10.1	66.5	74.3	72.2	78.9	70.4	81.7	88.7	92.1	62.1	71.0

Table 1. Results of Whole-body pose estimation on COCO-WholeBody [19, 48] V1.0 dataset. The teacher that guides DWPose-l and DWPose-m,s,t is RTMPose-x and RTMPose-l, respectively. "†" indicates multi-scale testing. Flip test is used.

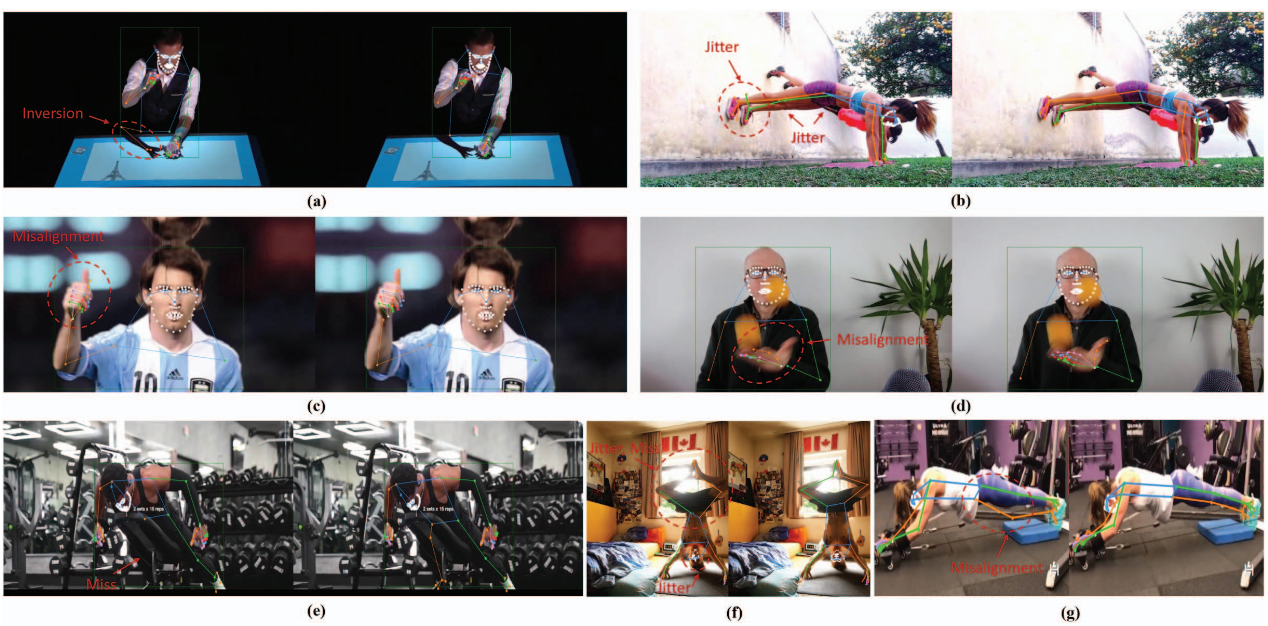

Figure 3. Qualitative comparisons of RTMPose-l (left) and DWPose-l (right). Best viewed in color with zoom-in for small parts.

RTMPose-x with fewer parameters and flops. DWPose-l also achieves the new state-of-the-art model for human whole-body pose estimation. With the proposed distillation TPD and more data, we provide a series of effective models with competitive accuracy.

Fig. 3 shows some qualitative comparisons of how our

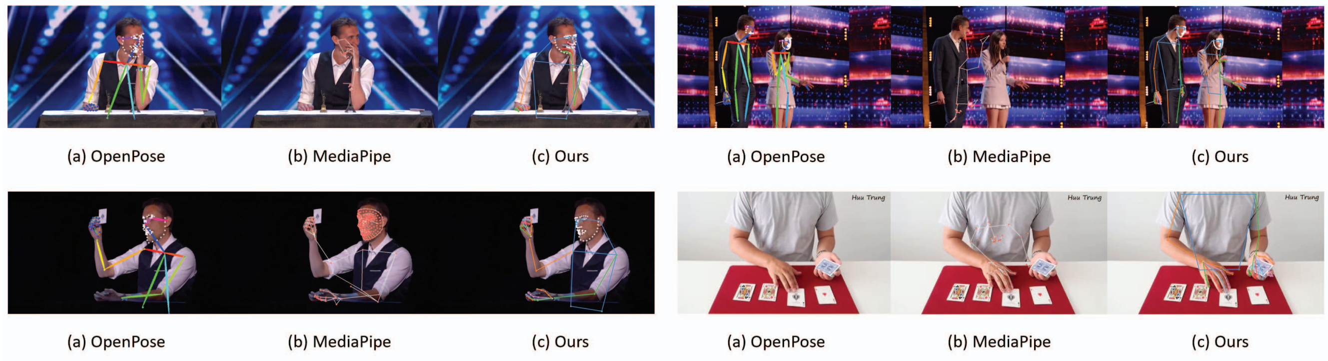

Figure 4. Qualitative comparisons with two popular whole-body pose estimators. (a) OpenPose; (b) MediaPipe; (c) Our DWPose-l.

Method	RTMPose* x-l		
UBody	-	✓	✓
TPD	-	-	✓
body	69.5	69.7	**70.4**
foot	65.8	65.5	**66.2**
face	83.3	84.1	**84.3**
hand	51.9	55.1	**56.6**
whole-body	61.1	62.1	**63.1**

Table 2. Ablation study of the Two-stages Pose Distillation (TPD) method and the UBody Dataset. The teacher and student models used are RTMPose-x and RTMPose-l, respectively.

	body	foot	face	hand	whole-body
RTM-l (256×192)	67.6	16.8	71.2	35.9	54.2
RTM-l*	69.3	24.3	71.3	49.6	59.8
DWPose-l*	**69.5**	**24.5**	**71.6**	**49.9**	**60.1**
RTM-l (384×288)	68.6	18.1	**73.8**	41.1	58.1
RTM-l*	69.4	24.7	72.4	54.6	62.9
DWPose-l*	**69.8**	**25.0**	73.4	**55.7**	**63.4**

Table 3. Results of evaluating the models on the UBody dataset. '*' indicates the models trained on both COCO and UBody datasets. The numbers are AP scores for two different input sizes.

distillation helps the students to perform better. TPD helps the model to predict more accurately, reduces false pose detection, and increases true pose detection, especially for the improvement of finger keypoint localization. We also compare our state-of-the-art model with two widely used models OpenPose [2] and MediaPipe [29, 62], as presented in Fig. 4. Our DWPose also surpasses the other two methods significantly, especially for the robustness of truncation, occlusion, and effectiveness of fine-grained localization. This enables our method to replace these popular methods to benefit corresponding downstream applications effectively.

5. Analysis

5.1. Effects of TPD Method and UBody Data

We explore the effects of the proposed distillation method TPD and the used UBody dataset of improving whole-body pose estimation in Tab. 2. The model achieves considerable AP improvements with an extra UBody dataset, especially for hand pose detection. For example, RTMPose-l achieves 62.1 whole-body AP and 55.1 Hand AP, which is 1.0 and 3.2 higher than the model trained just on COCO-Wholebody. Our distillation method TPD further boosts the model's performance, helping the model to get 63.1% whole AP. The results demonstrate the effectiveness of the TPD method and UBody dataset.

5.2. Performance on UBody

We first evaluate our method on the COCO WholeBody dataset, as we describe above. In this subsection, we evaluate the models on the UBody dataset, as shown in Tab. 3. We compare the models under two different input resolutions and report the corresponding AP of different human parts. The extra UBody data for training and our distillation method TPD are both helpful to the students, bringing them significant improvements under both input resolutions. Different from COCO, the gains that our TPD brings on UBody mainly focus on the face and hand. As for COCO, the performance on the body, foot, and hand all get significant improvements, but the gains on the face are limited, as shown in Tab. 2. The results on UBody also demonstrate the effectiveness of our distillation method TPD.

5.3. Effects of First and Second Stage Distillation

We propose the two-stage pose distillation (TPD), which includes the first and second stage distillation. To evaluate the impact of each distillation stage, we conduct experiments by using RTMPose-x to distill RTMPose-l on the mixed dataset, as presented in Tab. 4. Both two distillation stages are beneficial for the students, and their combination leads to further improvements in performance. When combining the first-stage and second-stage distillation together, we achieve 63.1 whole AP, which surpasses the performance achieved by using either distillation loss alone.

Method	RTMPose* x-l			
First-stage	-	✓	-	✓
Second-stage	-	-	✓	✓
body	69.7	**70.4**	69.7	**70.4**
foot	65.5	65.8	65.9	**66.2**
face	84.1	84.1	84.2	**84.3**
hand	55.1	56.4	55.4	**56.6**
whole-body	62.1	62.9	62.2	**63.1**

Table 4. Ablation study of the two distillation stages. The teacher and student are RTMPose-x and RTMPose-l. '*' denotes the model is trained on COCO + UBody.

	body	foot	face	hand	whole-body
RTMPose-m	69.1	64.8	81.8	49.8	60.0
RTMPose-m + S2	69.4	65.1	81.9	50.3	60.4
RTMPose-l	69.5	65.8	83.3	51.9	61.1
RTMPose-l + S2	69.6	66.1	83.2	52.3	61.3
RTMPose-m*	68.6	63.6	82.5	52.3	60.4
RTMPose-m* + S2	68.5	63.6	82.8	52.7	60.6
RTMPose-l*	69.7	65.5	84.1	55.1	62.1
RTMPose-l* + S2	69.7	65.9	84.2	55.4	62.2
RTMPose-x*	70.3	65.3	84.9	56.4	63.0
RTMPose-x* + S2	70.4	65.3	84.9	56.6	63.2

Table 5. The impact of the proposed head-aware self-KD in the second-stage distillation (S2) on existing estimator RTMPose. '*' denotes the model is trained on COCO + UBody. All results are reported with AP on COCO-WholeBody.

It's worth noting that the second-stage distillation just needs to fine-tune the head, which helps to save much training time. Interestingly, it helps the student to surpass the teacher RTMPose-x with 63.0% AP.

5.4. Second-stage Distillation for Trained Models

Our second-stage distillation is available not only for the models trained with our first-stage distillation but also for those trained without distillation. So it can be applied when there lacks a better and larger teacher. We can utilize the model itself as a teacher to improve it with a short training time. As shown in Tab. 5, we pick three different models and evaluate our second-stage distillation on COCO and the combination of COCO and UBody. For all settings, models with *S2* significantly improve, especially for the foot and hand. Compared with traditional distillation and self-KD, it saves much time in training the model from scratch and costs to obtain a better model.

5.5. Ablation Study of the First-stage Distillation

As we describe in Eq. 6, our first-stage distillation calculates the loss through the ground-truth label (GT), teacher's feature (Fea), and teacher's logits (Logit). Furthermore, we apply a weight-decay strategy (Decay) to further improve the student. In this subsection, we analyze the effects of every component by using RTMPose-l to distill RTMPose-m, as shown in Tab. 6. The knowledge from the feature brings the student 1.4% AP gains. When combing the distillation on the logit, the AP gains get to 1.6%. This proves that the knowledge from the feature and logit are both helpful and complementary to each other. Finally, the weight-decay strategy brings another 0.3% AP gains, helping the student to achieve 62.3% AP.

Interestingly, we try to drop the GT label and train the student just with the teacher's logit. The student achieves 60.9% AP, which is even 0.5% higher than the model trained with the GT label. This indicates we can label the new data through a teacher model instead of annotating manually, which can save much cost in time and manual efforts, and achieve a better model through such data for training. However, when combining the feature distillation together, the performance with the teacher's logit gets lower than that with the GT label. Thus, we adopt the GT, Fea, and Logit together for distillation.

GT	Fea	Logit	Decay	whole-body
✓	-	-	-	60.4
✓	✓	-	-	61.8
-	-	✓	-	60.9
-	✓	✓	-	61.4
✓	✓	✓	-	62.0
✓	✓	✓	✓	**62.3**

Table 6. Ablation study of the components of first-stage distillation. The teacher and student are RTMPose-l and RTMPose-m. The performance is the whole-body AP on COCO with GT boxes.

Logit	Mask	whole-body
✓	-	60.9
✓	✓	59.8

Table 7. Ablation study of the target weight mask. The teacher and student is RTMPose-l and RTMPose-m. The performance is the whole-body AP on COCO with GT boxes.

5.6. Target Mask for Logit-based Distillation

In our logit-based distillation, we deliberately omit the target weight mask W, which is employed to differentiate between visible and invisible keypoints, as shown in Eq. 3. We conducted an in-depth investigation into how this target mask affects the distillation process. As indicated in Tab. 7, it is evident that the presence of the target weight mask significantly hampers the distillation performance, resulting in a notable 1.1% drop in the student's performance. These results underscore the significance of the teacher's input for invisible keypoints, affirming its positive impact on the student's learning process.

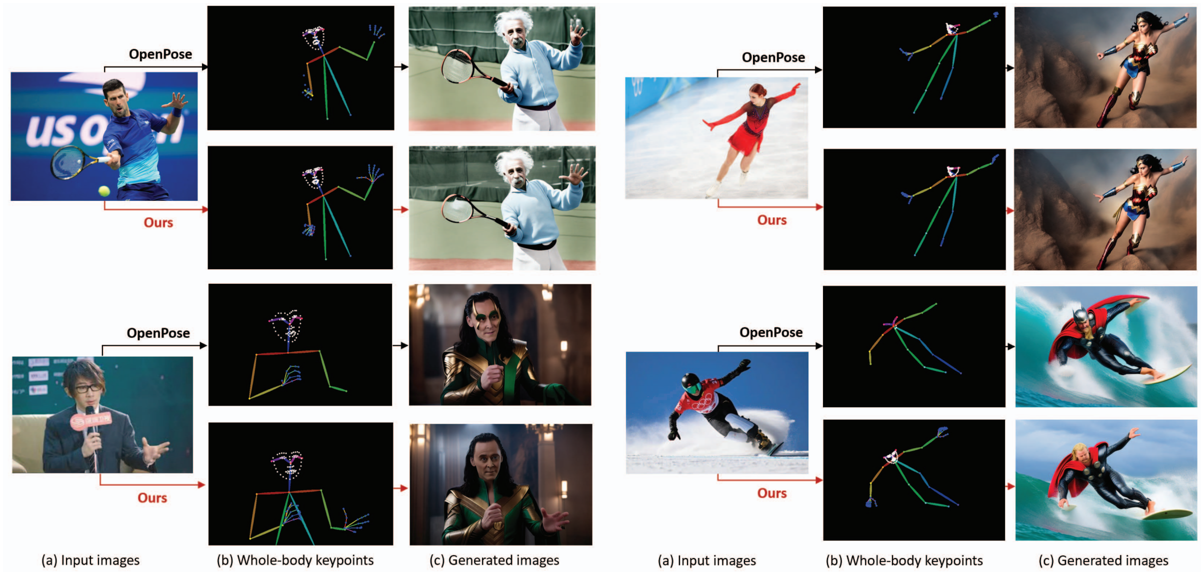

Figure 5. Visualization of skeleton-guided image generation results. The upper images depict the baseline using the pre-trained ControlNet v1.1 with the original estimator OpenPose. In contrast, the lower images showcase ControlNet with our DWPose-l as inputs. The prompts, random seeds, and other settings remain constant.

Num. of persons	1	2	3	4	5	6	7	8	9
OpenPose	5.78	6.28	7.55	8.90	9.16	10.4	15.2	16.86	18.91
Ours	0.068	0.070	0.077	0.082	0.084	0.088	0.094	0.098	0.108

Table 8. Comparison of pose estimation speed with OpenPose. The table displays the average time cost (in seconds) for inferring an image on an Nvidia RTX 3090 with varying numbers of persons present. We use YOLOX-l and DWPose-l to test the speed.

5.7. Better Pose, Better Image Generation

Recently, controllable image generation [12, 37, 39, 63, 33, 20] has witnessed significant advancements. For human image generation, precise skeleton information is crucial to guide the pose, particularly for whole-body skeletons. Mainstream techniques like ControlNet [63] often rely on OpenPose [2] due to its efficiency and user-friendly nature in generating human poses. However, OpenPose's performance, as shown in Tab. 1, reaches only 44.2% AP, which leaves room for improvement. Consequently, we aim to replace OpenPose with our DWPose to enhance ControlNet's image generation without the need for additional training. Utilizing a top-down approach, we first employ YOLO-X [11] to detect all individuals and then use our pose estimator to extract keypoints from the detection results, thus boosting the overall image generation process.

In Fig. 5, we employ ControlNet to visualize and compare the generated images using both OpenPose and our DWPose, demonstrating that a more precise and expressive skeleton leads to higher-quality image generation. Additionally, we present a comparison of inference speed with OpenPose in Tab. 8. Thanks to the efficient architecture of RTMPose [18], DWPose requires only about one percent of the time taken by OpenPose to infer the same image. Moreover, as the number of persons in the image increases, the runtime for OpenPose significantly increases. For a single person, the inference times for OpenPose and DWPose are 5.78 s and 0.068 s, respectively. However, when the number of persons reaches nine, the inference time for OpenPose triples, whereas the inference time for DWPose is only about 1.5 times longer.

6. Conclusion

In this paper, we aim to obtain both an efficient and effective model for human whole-body pose estimation. To this end, we apply distillation to the latest effective RTMPose. Accordingly, we first propose a Two-stage Pose Distillation to enhance the lightweight model's performance. Moreover, the second-stage distillation is available when a larger teacher lacks, and it only needs a short training time to obtain a better model. Then, we investigate the UBody dataset to further improve its performance, obtaining DWPose. Extensive experiments prove that our method is simple yet effective. We also explore the impact of a better pose estimator on the controllable image generation task.

Acknowledgement. This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), the SZSTC project Grant (JCYJ20190809172201639, WDZC20200820200655001), Shenzhen Key Laboratory (ZDSYS20210623092001004).

References

[1] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 1, 2, 5, 6, 8

[3] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 2, 5

[4] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022. 2

[5] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 4

[6] Bardia Doosti. Hand pose estimation: A survey. *arXiv preprint arXiv:1903.01013*, 2019. 2

[7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 1

[8] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16091–16101, 2023. 1

[9] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 5

[10] Maria-Paola Forte, Peter Kulits, Chun-Hao P Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J Kuchenbecker, and Michael J Black. Reconstructing signing avatars from video using linguistic priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12791–12801, 2023. 1

[11] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 8

[12] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*, 2023. 8

[13] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6982–6991, 2019. 2, 5

[14] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2

[15] Tianjin Huang, Lu Yin, Zhenyu Zhang, Li Shen, Meng Fang, Mykola Pechenizkiy, Zhangyang Wang, and Shiwei Liu. Are large kernels better teachers than transformers for convnets? *International Conference on Machine Learning*, 2023. 2

[16] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 2022. 2

[17] Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15996–16005, 2023. 2

[18] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 2, 3, 4, 8

[19] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214, 2020. 2, 3, 4, 5

[20] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. In *IEEE Conference on International Conference on Computer Vision*, 2023. 8

[21] Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4185, 2022. 2

[22] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pages 89–106, 2022. 2, 3

[23] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11740–11750, 2021. 2

[24] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *arXiv preprint arXiv:2307.00818*, 2023. 1

[25] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 1, 2, 4

[26] 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, 2014. 2, 4

[27] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022. 1

[28] Ruiping Liu, Kailun Yang, Alina Roitberg, Jiaming Zhang, Kunyu Peng, Huayao Liu, and Rainer Stiefelhagen. Transkd: Transformer knowledge distillation for efficient semantic segmentation. *arXiv preprint arXiv:2202.13393*, 2022. 2

[29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1, 2, 6

[30] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2

[31] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 2308–2317, 2022. 1

[32] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564, 2020. 2

[33] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 8

[34] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in Neural Information Processing Systems*, 30, 2017. 5

[35] 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. 2

[36] 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. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 8

[38] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 2

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015. 8

[40] 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. 2, 5

[41] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020. 4

[42] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pages 581–600, 2020. 1

[43] 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. 5

[44] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 5

[45] 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, 2020. 2, 3

[46] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127:115–142, 2019. 2

[47] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, pages 466–481, 2018. 2, 4, 5

[48] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 5

[49] Xianzhe Xu, Yiqi Jiang, Weihua Chen, Yilun Huang, Yuan Zhang, and Xiuyu Sun. Damo-yolo: A report on real-time object detection design. *arXiv preprint arXiv:2211.15444*, 2022. 2

[50] Xixia Xu, Qi Zou, Xue Lin, Yaping Huang, and Yi Tian. Integral knowledge distillation for multi-person pose estimation. *IEEE Signal Processing Letters*, 27:436–440, 2020. 2

[51] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose+: Vision transformer foundation model for generic body pose estimation. *arXiv preprint arXiv:2212.04246*, 2022. 5

[52] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2020. 2

[53] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[54] Zhendong Yang, Zhe Li, Yuan Gong, Tianke Zhang, Shanshan Lao, Chun Yuan, and Yu Li. Rethinking knowledge distillation via cross-entropy. *arXiv preprint arXiv:2208.10139*, 2022. 2

[55] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 2

[56] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European Conference on Computer Vision*, pages 53–69, 2022. 2

[57] Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. Vitkd: Practical guidelines for vit feature knowledge distillation. *arXiv preprint arXiv:2209.02432*, 2022. 2

[58] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. *arXiv preprint arXiv:2303.13005*, 2023. 2

[59] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 1

[60] Chong Yu, Tao Chen, Zhongxue Gan, and Jiayuan Fan. Boost vision transformer with gpu-friendly sparsity and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22658–22668, 2023. 2

[61] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 2

[62] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 1, 2, 6

[63] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 8

[64] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022. 2

[65] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *arXiv preprint arXiv:2305.04789*, 2023. 1

[66] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2020. 2

[67] Congcong Zhu, Xiaoqiang Li, Jide Li, and Songmin Dai. Improving robustness of facial landmark detection by defending against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11751–11760, 2021. 2

[68] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2

[69] Martin Zong, Zengyu Qiu, Xinzhu Ma, Kunlin Yang, Chunya Liu, Jun Hou, Shuai Yi, and Wanli Ouyang. Better teacher better student: Dynamic prior knowledge for knowledge distillation. In *The Eleventh International Conference on Learning Representations*, 2022. 2