MDPose: Real-Time Multi-Person Pose Estimation via Mixture Density Model

Anonymous authors
Paper under double-blind review

Abstract

One of the major challenges in multi-person pose estimation is instance-aware keypoint estimation. Previous methods address this problem by leveraging an off-the-shelf detector, heuristic post-grouping process or explicit instance identification process, hindering further improvements in inference speed which is an important factor for practical applications. From the statistical point of view, those additional processes for identifying instances are necessary to bypass learning the high-dimensional joint distribution of human keypoints, which is a critical factor for another major challenge, the occlusion scenario. In this work, we propose a novel framework of single-stage instance-aware pose estimation by modeling the joint distribution of human keypoints with a mixture density model, termed as MDPose. Our MDPose estimates the distribution of human keypoints' coordinates using a mixture density model with an instance-aware keypoint head consisting simply of 8 convolutional layers. It is trained by minimizing the *negative log-likelihood* of the ground truth keypoints. Also, we propose a simple yet effective training strategy, Random Keypoint Grouping (RKG), which significantly alleviates the underflow problem leading to successful learning of relations between keypoints. On OCHuman dataset, which consists of images with highly occluded people, our MDPose achieves state-of-the-art performance by successfully learning the high-dimensional joint distribution of human keypoints. Furthermore, our MDPose shows significant improvement in inference speed with a competitive accuracy on MS COCO, a widely-used human keypoint datasets, thanks to the proposed much simpler single-stage pipeline.

1 Introduction

Multi-person pose estimation is a classical computer vision task that aims to localize human keypoints in an image. As it is a fundamental computer vision problem leading to various practical applications such as action recognition, human-computer interaction and so on, it has been studied actively since the development of deep learning.

One of the major challenges in multi-person pose estimation is *instance-aware keypoint estimation* and many works have been studied to tackle this problem, which can be categorized into two major paradigms: top-down (Xiao et al., 2018; Sun et al., 2019; Li et al., 2021; Papandreou et al., 2017; Chen et al., 2018; Khirodkar et al., 2021) and bottom-up approaches (Varamesh & Tuytelaars, 2020; Zhou et al., 2019; Cao et al., 2017; Kreiss et al., 2019; Cheng et al., 2020; Geng et al., 2021; Newell et al., 2017). As shown in Figure. 1(a) and (b), the top-down method exploits an off-the-shelf detector and the bottom-up method performs a post-grouping process for a common goal of instance specification. However, there exist some bottlenecks toward the efficient instance-aware keypoint estimation. Since the top-down method is a two-stage method which detects a person then localizes its keypoints one by one, the more the number of people in an image, the slower the inference speed. In the case of the bottom-up method, it depends on a post-grouping process, which is usually heuristic and takes additional time for keypoint refinement for the instance-aware keypoint estimation.

Recently, there have been approaches to tackle the aforementioned weaknesses for instance-aware keypoint estimation (Tian et al., 2019a; Mao et al., 2021), as shown in Figure 1(c). Mao et al. (2021) proposed FCPose, a single-stage instance-aware framework based on FCOS detector (Tian et al., 2019b), equipped with a dynamic keypoint head consisting of instance-specific weights. Since it

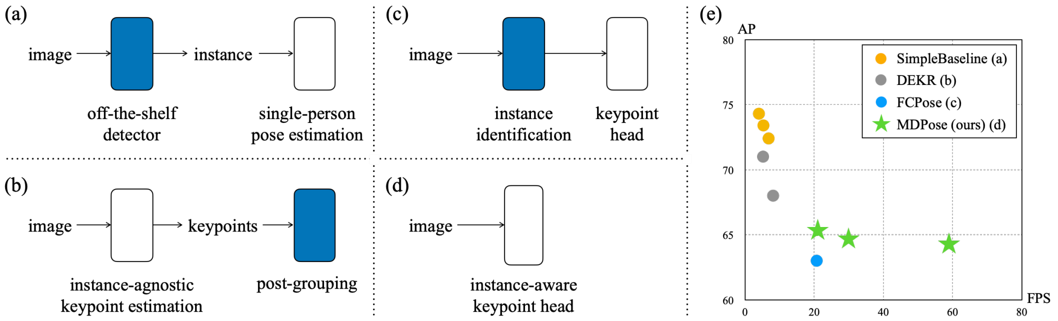


Figure 1: Illustration of multi-person pose estimation frameworks: (a) Top-down, (b) Bottom-up, (c) previous Single-stage Instance-aware, (d) Ours and (e) Speed-accuracy trade-off. The colored boxes indicate the process for identifying instances, which we successfully removed by proposing a mixture-model-based architecture. Details for (e) are provided in Table 3.

leverages the capacity of FCOS detector and is a one-stage method at the same time, it can achieve a reasonably high accuracy at a relatively fast inference speed. However, it still relies on the detector's performance for generating instance weights and such instance identification process hinders further improvement in the inference speed.

In this paper, we propose a novel multi-person pose estimation framework using a mixture model. There has been a line of research utilizing the mixture model in various pose estimation tasks (Li & Lee, 2019; Prokudin et al., 2018; Ye & Kim, 2018; Varamesh & Tuytelaars, 2020). Among them, MDN$_3$ (Varamesh & Tuytelaars, 2020) showed the potential in the multi-person pose estimation task by modeling the mixture model with a person's viewpoint as a dominant factor. However, it lags behind other state-of-the-art methods in terms of accuracy and inference speed.

Inspired by MDOD (Yoo et al., 2021), which showed competitive performance with a mixture-model-based architecture in object detection, we propose a simple architecture modeling joint distribution of human keypoints with a mixture model, coined as MDPose. From the statistical point of view, previous methods need to implement an additional instance identification process to bypass learning high-dimensional joint distribution of human keypoints' coordinates, since the numerical underflow problem usually occurs during the training process due to the curse of dimensionality. However, unless the high-dimensional distribution is considered sufficiently, the performance degradation is inevitable under the condition of severe occlusion. To tackle this problem, we propose *Random Keypoint Grouping* (RKG) which learns the joint distribution of continuously changing subsets of keypoints at every iteration. It alleviates the underflow problem efficiently and leads to the successful learning of relations between keypoints in the high-dimensional space, which increases the capacity for distinguishing multiple occluded persons. Furthermore, since a mixture component corresponds to a person, we can perform instance-aware keypoint estimation without any additional instance identification process, as shown in Figure 1(d). As a result, we could achieve competitive performances with a simple instance-aware keypoint head consisting of only 8-convolutional layer enabling real-time applications. Additionally, unlike Sun et al. (2019); Xiao et al. (2018); Cheng et al. (2020); Cao et al. (2017); He et al. (2017); Geng et al. (2021); Newell et al. (2017), MDPose does not need likelihood heatmap during training which requires burdensome computational cost and storage. In short, MDPose shows strong potential for practical applications with regard to both training and inference time as well as an occlusion condition.

Our MDPose performs instance-aware keypoint estimation without bells and whistles through a mixture model framework. Our RKG makes it possible to learn high-dimensional joint distribution of human keypoints' coordinates, eliminating additional instance identification processes. Specifically, on the OCHuman (Zhang et al., 2019) validation and test set consisting of images with heavily occluded persons, our MDPose achieves state-of-the-art performance with **43.5** mAP^{kp} and **42.7** mAP^{kp}, respectively, by successfully learning human keypoint representation in a high-dimensional space. Furthermore, on the COCO keypoint validation set (Lin et al., 2014), our MDPose achieves **64.6** mAP^{kp} at the speed of **29.8** FPS with a ResNet-50 backbone (He et al., 2016), which outperforms other state-of-the-art methods by a large margin in inference speed (see Figure 1(e)).

2 Related works

Multi-person pose estimation. One of the major challenges in multi-person pose estimation is to correctly estimate keypoints per each person, i.e. instance-aware keypoint estimation. Many studies have been done to address this problem which can be classified into two paradigms: top-down and bottom-up approaches. The top-down approach (Papandreou et al., 2017; Chen et al., 2018; Xiao et al., 2018; Sun et al., 2019; Li et al., 2021; Khirodkar et al., 2021) leverages an off-the-shelf detector to localize an instance and performs a single-person pose estimation. While it can achieve high accuracy, its inference speed is much slower than bottom-up approaches, especially for an image with a large number of people. On the other hand, the bottom-up approach (Newell et al., 2017; Zhou et al., 2019; Cao et al., 2017; Kreiss et al., 2019; Cheng et al., 2020; Geng et al., 2021; Xue et al., 2022) performs instance-agnostic keypoint estimation and assigns them to each instance through a post-grouping process. It shows more robust and faster inference speed than top-down approaches. However, the post-grouping process is usually heuristic and complicated with many hyperparameters.

Single-stage instance-aware pose estimation. Recently, there have been single-stage instance-aware approaches to tackle the aforementioned drawbacks of existing frameworks (Tian et al., 2019a; Mao et al., 2021). Among them, Mao et al. (2021) proposed end-to-end trainable FCPose which performs instance-aware keypoint estimation by a dynamic keypoint head consisting of instance-specific weights generated by FCOS detector (Tian et al., 2019b). As a result, it achieves competitive accuracy and inference speed while eliminating heuristic post-grouping process. However, it still depends on the performance of the object detector for identifying instances and the instance-specific weight generation process remains as a bottleneck for further improvement of inference speed.

Occluded pose estimation. There are various approaches (Jin et al., 2020; Khirodkar et al., 2021; Li et al., 2019; Qiu et al., 2020; Zhang et al., 2019) to improve performance in occluded human pose estimation, which is another major challenge. Jin et al. (2020) proposed a hierarchical graph grouping method to learn relationship between keypoints in the bottom-up style. Among the top-down methods, Khirodkar et al. (2021) introduced a Multi-Instance Modulation Block which adjusts feature responses to distinguish multiple instances in a given bounding box. Although they improve performance in the occlusion condition by specifically devised methods or architectures, they still lack enough consideration for learning the high-dimensional distribution of keypoints, which is a fundamental challenge in the multi-person pose estimation.

Mixture model in multi-person pose estimation. There has been a line of research using mixture models in various computer vision tasks (Li & Lee, 2019; Prokudin et al., 2018; Ye & Kim, 2018; Varamesh & Tuytelaars, 2020; Yoo et al., 2021). Among them, Varamesh & Tuytelaars (2020) introduced a mixture density network (Bishop, 1994) to a CenterNet-based (Zhou et al., 2019) architecture and showed the possibility of using the mixture model in multi-person pose estimation. However, it ends up in just playing a role as an auxiliary factor while utilizing all the loss terms of CenterNet for training, and lags behind other state-of-the-art methods. We propose a novel single-stage instance-aware keypoint estimation framework using a mixture density model, which is coined as MDPose. It achieves state-of-the-art performance in a heavy occlusion condition and enables real-time estimation without any explicit instance identification process as shown in Fig. 1(d).

3 Method

In this work, we propose a novel framework for learning the joint distribution of human keypoints using a mixture model, leading to eliminating explicit instance identification processes and boosting the capacity of distinguishing occluded persons. Our MDPose is modeled with a mixture distribution so that the mixture component corresponds to a person, i.e. one-to-one matching between mixture components and persons, resulting in instance-aware keypoint estimation without bells and whistles. Since it depends on neither an off-the-shelf detector nor a post-grouping process, it can achieve a much simpler pipeline with an accelerated speed than previous methods.

First, we will describe the mixture model and our problem formulation in Sec. 3.1 and propose a new architecture and describe it in detail in Sec. 3.2. After that, we will explain the *Random*

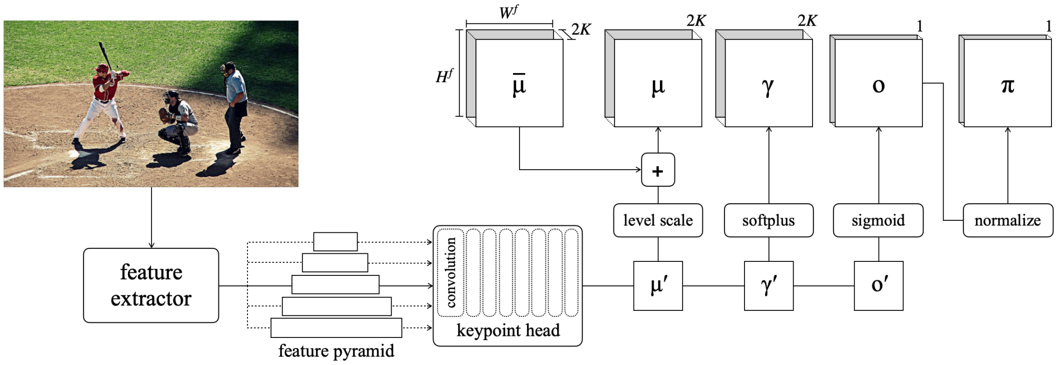


Figure 2: The overall architecture of MDPose. The parameters of mixture model (μ, γ, o and π) are obtained from a keypoint head consisting of 8 convolutional layers. The mixture components are located along the spatial axis, i.e. the number of mixture components in a feature map is $H^f \times W^f$.

Keypoint Grouping (RKG) strategy for learning the high-dimensional joint distribution and our final loss function in Sec. 3.3. Finally, an inference phase will be described in Sec. 3.4.

3.1 Mixture model

In an image X, there exists a ground truth for each of N persons, $k^{gt} = \{k_1^{gt}, \dots, k_N^{gt}\}$, and i-th ground truth k_i^{gt} contains the keypoint coordinates $k_i^{gt} = \{k_{i,1,x}, k_{i,1,y}, \dots, k_{i,K,x}, k_{i,K,y}\}$, where K denotes the number of keypoints. Our MDPose estimates the distribution of keypoint locations k_i on an image X with a mixture model. Based on the design of the mixture model for object detection in Yoo et al. (2021), we develop and modify the architecture for the multi-person pose estimation task. Our mixture model is formed by a weighted combination of component distributions, which we set as a Laplace distribution. Although the Laplace distribution has a similar shape with the Gaussian and the Cauchy distribution, its tails fall off **more rapidly** than the Cauchy but **less sharply** than the Gaussian. We empirically found that the Laplace distribution is more suitable for the multi-person pose estimation task than the Gaussian and Cauchy. Related experimental results are provided in the appendix. Following Yoo et al. (2021), every element of k_i is assumed to be independent[1] of each other to keep the mixture model from being over-complicated. Therefore, the probability density function (pdf) of Laplace distribution is defined as,

$$\mathcal{F}(k_i; \mu, \gamma) = \prod_{j=1}^{K} \prod_{d \in D} \mathcal{F}(k_{i,j,d}; \mu_{j,d}, \gamma_{j,d}) = \prod_{j=1}^{K} \prod_{d \in D} \frac{1}{2\gamma_{j,d}} \exp\left(-\frac{|k_{i,j,d} - \mu_{j,d}|}{\gamma_{j,d}}\right) \tag{1}$$

with a set of keypoint coordinates $D = \{x, y\}$, where j and $\mathcal{F}$ are the keypoint index and the Laplacian pdf, respectively. As a result, the $2K$-dimensional Laplace represents the distribution of human keypoints coordinates and the pdf of our mixture model is as follows:

$$p(k_i^{gt}|X) = \sum_{m=1}^{M} \pi_m \mathcal{F}(k_i; \mu_m, \gamma_m), \tag{2}$$

where the m denotes the index of M mixture components.

3.2 Architecture

Figure 2 demonstrates the overall architecture of our MDPose. The feature maps are forwarded into the keypoint head to obtain intermediate outputs: μ', γ', and o'. The final outputs μ, γ, o, and π are obtained from intermediate outputs as parameters of our mixture model. The mixture components are represented at each position of the cells on the feature map, i.e. located along the spatial axis.

[1] Although each element of a mixture component is independent of others, they are jointly dependent in the overall joint distribution.

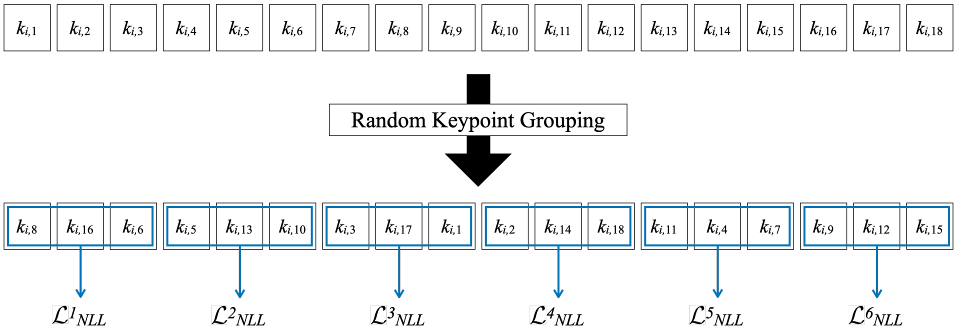

$k_{i,1}$ $k_{i,2}$ $k_{i,3}$ $k_{i,4}$ $k_{i,5}$ $k_{i,6}$ $k_{i,7}$ $k_{i,8}$ $k_{i,9}$ $k_{i,10}$ $k_{i,11}$ $k_{i,12}$ $k_{i,13}$ $k_{i,14}$ $k_{i,15}$ $k_{i,16}$ $k_{i,17}$ $k_{i,18}$
Random Keypoint Grouping
$k_{i,8}$ $k_{i,16}$ $k_{i,6}$ $k_{i,5}$ $k_{i,13}$ $k_{i,10}$ $k_{i,3}$ $k_{i,17}$ $k_{i,1}$ $k_{i,2}$ $k_{i,14}$ $k_{i,18}$ $k_{i,11}$ $k_{i,4}$ $k_{i,7}$ $k_{i,9}$ $k_{i,12}$ $k_{i,15}$
$\mathcal{L}^1{}_{NLL}$
$\mathcal{L}^2{}_{NLL}$
$\mathcal{L}^3{}_{NLL}$
$\mathcal{L}^4{}_{NLL}$
$\mathcal{L}^5{}_{NLL}$
$\mathcal{L}^6{}_{NLL}$

Figure 3: Illustration of RKG at an iteration, with $K_g = 3$ and $N_g = 6$. $k_{i,j}$ is a human keypoint, where i and j denote a person in an image and a keypoint index, respectively. For the simplicity of grouping, we set the center coordinate of the bounding box as $k_{i,18}$.

The mean μ is derived from $\mu' \in \mathbb{R}^{H^f \times W^f \times 2K}$, where H^f and W^f indicate the height and width of a feature map in the feature pyramid, respectively, and note that the number of mixture components in a feature map is $H^f \times W^f$. First, following the implementation of Yoo et al. (2021), μ' is scaled by a factor of $s = 2^{l-5}$, where $l \in \{1, \cdots, 5\}$ denotes the level of feature map in the feature pyramid. Then, the scaled μ' is added to $\bar{\mu} \in \mathbb{R}^{H^f \times W^f \times 2K}$ which is the default coordinates uniformly distributed in a grid pattern over the entire feature map. In short, the final location parameter μ is obtained as follows: $\mu = \bar{\mu} + s\mu'$. We can obtain the positive scale parameter $\gamma \in \mathbb{R}^{H^f \times W^f \times 2K}$ through softplus (Dugas et al., 2000) activation function from γ'. The foreground probability $o \in \mathbb{R}^{H^f \times W^f \times 1}$ is calculated by applying the sigmoid function to o'. Following Yoo et al. (2022), we use the normalized foreground probability as π: $\pi_m = o_m / \sum_{n=1}^{M} o_n$. Since the mixture components in a foreground area are likely to have higher π, we can consider π as the normalized foreground probability so that $\sum_m^M \pi_m = 1$.

The keypoint head of MDPose consists of eight 3x3 kernel convolutional layers with Swish (Ramachandran et al., 2017) activation function except the last layer. The 5-level Feature Pyramid Network (Lin et al., 2017) is used as our feature extractor. Since we estimate a mixture distribution from all-level feature maps, the total number of mixture components is equal to the summation of the number of mixture components in each level of feature map: $M = \sum_{l=1}^{5} (H_l^f \times W_l^f)$.

3.3 Training

Our MDPose is trained to maximize the likelihood of k^{gt} for an input image X. Therefore, we can simply define the loss function for minimizing the *negative log-likelihood* (NLL) of k^{gt} as follows:

$$\mathcal{L}_{NLL} = -\log p(k^{gt}|X) = -\log \prod_{i=1}^{N} p(k_i^{gt}|X) = -\sum_{i=1}^{N} \log \sum_{m=1}^{M} \pi_m \mathcal{F}(k_i; \mu_m, \gamma_m). \tag{3}$$

Although the foreground probability o is not used to calculate $\mathcal{L}_{NLL}$, it is trained through the mixture coefficient π, i.e. the probability of a mixture component (Yoo et al., 2022).

In the training using (3), the curse of dimensionality arises due to the high-dimensional joint distribution of human keypoints, e.g. 34 dimension in the case of 17 keypoints in COCO keypoint dataset (Lin et al., 2014), leading to a severe underflow problem. As a result, it is extremely hard to compute $\mathcal{L}_{NLL}$ via a $2K$-dimensional joint distribution in the multi-person pose estimation task.

Random keypoint grouping (RKG). To tackle this problem, we propose RKG. As illustrated in Figure 3, we shuffle and split K keypoints into N_g groups, each consisting of K_g keypoints, where N_g and K_g denote the number of groups and the number of keypoints in a group, respectively, i.e. $K_g \times N_g = K$. As a result, we can notate a set of keypoints' indices in a group g as I_g and reformulate (1) using a group of keypoints as follows:

$$\mathcal{F}(k_i^g; \mu^g, \gamma^g) = \prod_{j \in I_g} \prod_{d \in D} \mathcal{F}(k_{i,j,d}; \mu_{j,d}, \gamma_{j,d}), \tag{4}$$

where the superscript g indicates the index of the group. Therefore, we can alleviate the underflow problem with $2K_g$-dimensional joint distribution, whose dimension is lower than the original $2K$ dimension if $K_g < K$. Our final loss function with RKG is defined as follows:

$$\mathcal{L}_{GroupNLL} = \frac{\sum_{g=1}^{N_g} \mathcal{L}^{g}_{NLL}}{N_g} = -\frac{1}{N_g}\sum_{i=1}^{N}\sum_{g=1}^{N_g}\log\sum_{m=1}^{M}\pi_m\mathcal{F}(k_i^g;\mu_m^g,\gamma_m^g). \tag{5}$$

Note that RKG is used only for the training process and the combination of keypoints for a group changes at every iteration. As shown in (5), the RKG amounts to factorizing the original joint distribution of $2K$ dimension into N_g marginal distributions of $2K_g$ dimension. Although each keypoints group is estimated independently at each iteration, the keypoints end up being dependent on each other through the whole training process due to RKG, which keeps shuffling and grouping keypoints randomly. As a result, MDPose is able to learn the relations between keypoints without any heuristic grouping process. To ease the grouping scheme for COCO keypoint dataset (Lin et al., 2014) labeled with 17 keypoints, we use the coordinates of bounding box center of k_i^{gt} as an auxiliary keypoint only for training, which is denoted as $k_{i,18}$ in Figure 3.

3.4 Inference

In the inference phase, a mixture component of our MDPose corresponds to an instance, i.e. a person in the multi-person pose estimation task. Therefore, MDPose is able to perform an instance-aware keypoint estimation without bells and whistles. μ and o are used as the estimated keypoint coordinates and confidence scores, respectively. Note that we do not use μ of the bbox center coordinates for inference. Our final predictions are obtained by removing duplicate estimations using non-maximum suppression (NMS), which is applied to pseudo-bboxes, each of which consists of the minimum and the maximum coordinates among keypoints as the left-top and the bottom-right coordinates, respectively.

4 Experiments

4.1 Experimental details

Dataset. We evaluate MDPose on the widely-used human keypoint dataset, MS COCO (Lin et al., 2014), consisting of 200K images including 250K person instances labeled with 17 keypoints. Following the standard protocol, we split the dataset into 57K images for training, 5K images for validation, and 20K images for test-dev set. We adopt the *average precision* (AP) based on the *object keypoint similarity* (OKS) as our evaluation metric. We conduct the analysis for our MDPose on the validation set and compare with other state-of-the-art methods on the test-dev set. Furthermore, we evaluate MDPose on OCHuman (Zhang et al., 2019), which is a *testing-only* dataset focusing on the heavy occlusion scenarios. It consists of 4,731 images with 8,110 person instances labeled with 17 keypoints like MS COCO. While less than 1% of person instances have occlusions with maxIoU $\geq$ 0.5 in MS COCO, all instances have occlusions with maxIoU $\geq$ 0.5 and 32% of them are more challenging with maxIoU $\geq$ 0.75 in OCHuman. Following Zhang et al. (2019), we use only MS COCO train set for training and evaluate on OCHuman validation and test set.

Training. As mentioned in 3.1, we represent the distribution of keypoint coordinates as a Laplace distribution. We set $K_g = 3$ and $N_g = 6$ for RKG as our default setting. We conduct experiments with different backbones including ResNet-50, ResNet-101 (He et al., 2016) and DLA-34 (Yu et al., 2018), which is especially for further improvement of inference speed. All backbones are pretrained with ImageNet (Deng et al., 2009) and FPN (Lin et al., 2017) is used as the feature extractor. For data augmentation, we apply random rotation in [-30°, 30°], expand, random crop in [0.3, 1.0] (relative range) and random flip. Unless specified, the input image is resized to 320×320 for the analysis of the RKG and mixture distributions or 896×896 for the analysis of inference speed and occluded pose estimation and comparison with other methods. Following Yoo et al. (2021), MDPose is trained by SGD with a weight decay of 5e-5 and gradient clipping with an L2 norm of 7.0. The batch size is 32 and the synchronized batch normalization (Peng et al., 2018) is used for a consistent learning behavior over different numbers of GPUs. The initial learning rate is set to 0.01 which is reduced by a factor of 10 at the 180K and 240K iteration in the training schedule of total 270K iterations.

Table 1: The performance according to the number of keypoints per group. K_g and N_g denote the number of keypoints per group and the total number of groups, respectively.

K_g	N_g	AP^{kp}	AP$^{kp}_{50}$	AP$^{kp}_{75}$	AP$^{kp}_{M}$	AP$^{kp}_{L}$
1	18	47.7	76.7	50.2	37.6	61.7
2	9	51.2	79.6	54.1	41.3	**64.9**
3	**6**	**51.5**	**80.4**	**55.1**	**42.0**	64.7
6	3	51.0	80.1	54.5	41.2	64.3
9	2	49.8	78.9	53.4	40.5	62.8
18	1	NaN	NaN	NaN	NaN	NaN

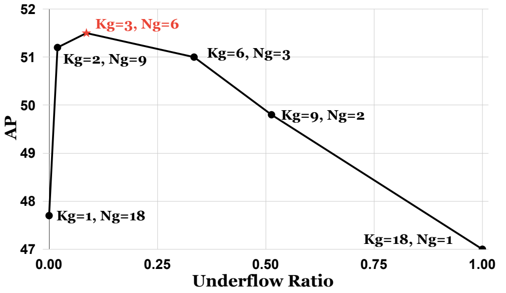


Figure 4: The trade-off between accuracy and ratio of underflowed components.

Inference. For inference, we use the same size of an image as in the training phase. The mixture components with low confidence scores in o are filtered out and NMS is applied for removing duplicate estimations. We set thresholds of o and NMS as 1e-4 and 0.7, respectively. Note that our model does not have any explicit process for identifying instance, such as post-grouping, weight generation and so on.

4.2 Analysis of RKG

The number of keypoints per group. We conducted an analysis for the number of keypoints per group, K_g. Since the number of groups, N_g, is determined according to K_g, i.e. $K_g \times N_g = K$, the more the number of keypoints in a group, the higher the joint distribution's dimension is.

Table 2: Randomness of grouping strategy. Non-random indicates the heuristic grouping method which predefines the keypoints per group based on the relations of human body joints.

Randomness	AP^{kp}	AP$^{kp}_{50}$	AP$^{kp}_{75}$	AP$^{kp}_{M}$	AP$^{kp}_{L}$
Non-random	39.5	69.9	40.0	32.4	49.7
Random	**51.5**	**80.4**	**55.1**	**42.0**	**64.7**

In Table 1, it shows the best performance of 51.5 AP^{kp} with RKG of $K_g = 3$, which we set as our default setting. Figure 4 shows the trade-off between the accuracy and the ratio of numerically underflowed components. When we apply RKG of $K_g = 1$ or 2, the performance is inferior to RKG of $K_g = 3$ despite the lower underflow ratio since our MDPose with RKG of high K_g can learn the relations of keypoints more efficiently by modeling the joint distribution with more keypoints. In particular, although there is no underflow problem due to the low dimension of joint distribution with RKG of $K_g = 1$, it cannot learn the relations of keypoints sufficiently during the training process, leading to notably lower AP^{kp} than RKG of $K_g = 2$ and 3 as shown in Table 1.

However, with RKG of more than $K_g = 3$, our MDPose suffers from the underflow problem as K_g increases, and the performance is rather lower than that with RKG of $K_g = 3$. As expected, with RKG of $K_g = 18$, i.e. with only one group, the original joint distribution is impossible to learn, resulting in NaN in Table 1. It is due to the severe underflow problem caused by the curse of dimensionality, i.e. the underflow ratio is 1.0 as shown in Figure 4.

Randomness in the grouping. Table 2 compares RKG with non-random grouping, which forms a group heuristically based on the relations of human body joints, i.e. $N_g = 6$ groups of *left arm*, *left leg*, *right arm*, *right leg*, *eyes and nose*, and *ears and bbox center*, each consisting of $K_g = 3$ keypoints. In comparison to the MDPose with non-random grouping, RKG improves the performance significantly from 39.5 AP^{kp} to 51.5 AP^{kp}. While non-random grouping learns only the joint distributions of pre-defined adjacent keypoint groups, RKG enables learning of the overall joint distributions of every non-adjacent keypoints through the whole training process by randomly grouping at every iteration. The comparison through qualitative results is provided in the appendix.

4.3 Analysis of the inference speed

Table 3 and Figure 1(e) present the comparison with other methods on COCO validation set. The FPS is measured on a single NVIDIA TITAN RTX. Our MDPose achieves 64.6 AP^{kp} and 29.8 FPS

Table 3: Inference speed comparison with other methods on COCO val set.

Method	Backbone	AP^{kp}	FPS
CenterNet (Zhou et al., 2019)	Hourglass	64.0	6.8
DEKR (Geng et al., 2021)	HRNet-W32	68.0	8.1
	HRNet-W48	71.0	5.2
FCPose (Mao et al., 2021)	ResNet-50	63.0	20.7
SimpleBaseline (Xiao et al., 2018)	ResNet-50	72.4	6.8
	ResNet-101	73.4	5.3
	ResNet-152	**74.3**	4.0
PifPaf (Kreiss et al., 2019)	ResNet-152	67.4	4.7
MDPose (Ours)	ResNet-50	64.6	29.8
	ResNet-101	65.2	20.8
	DLA-34	64.2	**58.9**

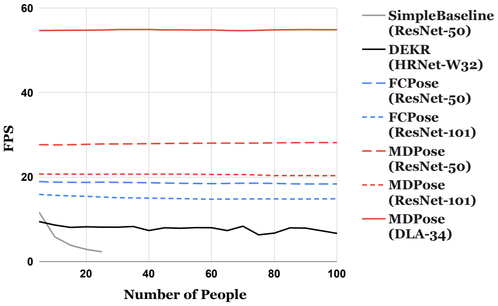
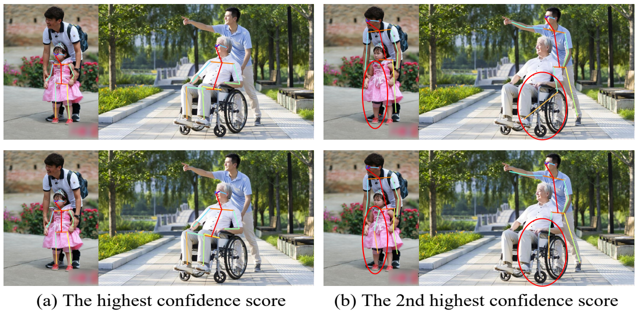


Figure 5: Inference speed by the number of people in an image.


(a) The highest confidence score

(b) The 2nd highest confidence score

Figure 6: Comparison between FCPose and MDPose in the occlusion scenario. FCPose and MDPose are shown on the 1st and 2nd row, respectively. The red circles in (b) show the differences of the estimated results for occluded keypoints between FCPose and MDPose.

with ResNet-50 backbone, which is comparable or superior to others especially in inference speed. It is 44%-faster than FCPose with an identical backbone, which is a single-stage instance-aware method enabling real-time application. Furthermore, we implement MDPose with DLA-34 as a backbone to further boost the inference speed. Following Tian et al. (2019b), we adopt the 3-level FPN and a training schedule of 360K iterations with learning rate decay by a factor of 10 at 300K and 340K iteration. The input image is resized to 736x736 for both training and inference. We can achieve about 3x-faster inference speed compared to FCPose (ResNet-50), still showing higher accuracy of 64.2 AP^{kp}.

Figure 5 illustrates the inference speed by the number of instances in an image. Our MDPose shows the robust inference speed, regardless of the number of people, even faster than FCPose. Furthermore, our MDPose with a heavier backbone ResNet-101 surpasses FCPose with ResNet-50 regarding the inference speed. It shows a strong potential of MDPose for the practical application enabling real-time multi-person pose estimation.

4.4 Analysis of the occluded pose estimation

Figure 6 shows comparison between FCPose (1st-row), a representative single-stage instance aware method, and MDPose (2nd-row) under the occlusion scenario. Figure 6(a) and (b) are the estimation results of person instances with the highest and 2nd highest confidence score, respectively.

As shown in the Figure 6(a), both of FCPose and MDPose estimate the keypoints of a person in the front successfully. However, for the person occluded by the other one, there exist two major drawbacks in FCPose. As demonstrated in the red circles in the 1st-row of Figure 6(b), FCPose misses a keypoint occluded by the other instance or confuses it with that of the other instance. As a result, it is not able to construct a proper form of human pose. Compared to FCPose, our MDPose estimates the occluded keypoints much more robustly by successfully learning the high-dimensional joint distribution of keypoints.

4.5 Comparison with state-of-the-art methods

OCHuman. Table 4 compares our MDPose with other state-of-the-art methods on OCHuman validation and test set. Note that we do not train our MDPose with OCHuman train set, but with

Table 4: Comparisons with SOTA methods on OCHuman val/test set. The evaluation metric is AP^{kp}.

Method	Backbone	Val.	Test
Top-down			
RMPE (Fang et al., 2017)	Hourglass	38.8	30.7
HRNet (Sun et al., 2019)	HRNet-W48	37.8	37.2
SimpleBaseline (Xiao et al., 2018)	ResNet-50	37.8	30.4
	ResNet-152	41.0	33.3
MIPNet (Khirodkar et al., 2021)	ResNet-101	32.8	35.0
	HRNet-W48	**42.0**	**42.5**
Bottom-up			
AE (Newell et al., 2017)	Hourglass	32.1	29.5
HGG (Jin et al., 2020)	Hourglass	35.6	34.8
DEKR (Geng et al., 2021)	HRNet-W32	37.9	36.5
	HRNet-W48	38.8	38.2
LOGO-CAP (Xue et al., 2022)	HRNet-W32	39.0	38.1
	HRNet-W48	**41.2**	**40.4**
Single-stage Instance-aware			
FCPose (Mao et al., 2021)	ResNet-50	32.4	31.7
	ResNet-101	33.3	33.4
MDPose (Ours)	ResNet-50	40.4	39.9
	ResNet-101	**43.5**	**42.7**

Table 5: Comparisons with SOTA methods on COCO test-dev set. We measure the inference speed of other methods on the identical hardware if possible. † denotes flipping in test time.

Method	Backbone	AP^{kp}	AP$^{kp}_{50}$	AP$^{kp}_{75}$	AP$^{kp}_{M}$	AP$^{kp}_{L}$	FPS
Top-down							
SimpleBaseline† (Xiao et al., 2018)	ResNet-152	73.7	91.9	81.1	70.3	80.0	2.3
HRNet† (Sun et al., 2019)	HRNet-W32	74.9	92.5	82.8	71.3	80.9	**3.0**
	HRNet-W48	75.5	**92.5**	**83.3**	71.9	**81.5**	2.0
RLE† (Li et al., 2021)	ResNet-152	74.2	91.5	81.9	71.2	79.3	-
	HRNet-W48	**75.7**	92.3	82.9	**72.3**	81.3	-
Bottom-up							
CMU-Pose (Cao et al., 2017)	VGG-19	61.8	84.9	67.5	57.1	68.2	**13.5**
MDN$_3^†$ (Varamesh & Tuytelaars, 2020)	Hourglass	62.9	85.1	69.4	58.8	71.4	7.0
CenterNet† (Zhou et al., 2019)	Hourglass	63.0	86.8	69.6	58.9	70.4	-
PifPaf (Kreiss et al., 2019)	ResNet-152	66.7	87.8	73.6	62.4	72.9	-
HigherHRNet† (Cheng et al., 2020)	HRNet-W32	66.4	87.5	72.8	61.2	74.2	2.5
	HRNet-W48	68.4	88.2	75.1	64.4	74.2	1.7
DEKR† (Geng et al., 2021)	HRNet-W32	67.3	87.9	74.1	61.5	76.1	8.5
	HRNet-W48	**70.0**	**89.4**	**77.3**	**65.7**	**76.9**	5.2
Single-stage Instance-aware							
DirectPose (Tian et al., 2019a)	ResNet-50	62.2	86.4	68.2	56.7	69.8	13.5
FCPose (Mao et al., 2021)	ResNet-50	64.3	87.3	71.0	61.6	70.5	20.3
	ResNet-101	**65.6**	87.9	72.6	**62.1**	**72.3**	15.5
MDPose (Ours)	ResNet-50	64.0	88.8	71.6	59.7	70.5	**28.7**
	ResNet-101	65.0	**88.9**	**72.8**	60.6	71.4	20.5

only MS COCO train set. Our MDPose outperforms other methods without bells and whistles due to the human keypoint representations successfully learned in the high-dimensional space by our mixture model with RKG. Compared to FCPose (ResNet-101), a state-of-the-art single-stage instance-aware method, our MDPose (ResNet-101) shows much better performance by a significant margin of **+10.2**%p AP^{kp} and **+9.3**%p AP^{kp} on the validation and test set, respectively. Furthermore, our MDPose (ResNet-101) even outperforms MIPNet (HRNet-W48), which was devised with more emphasis on the occlusion scenarios, by **+1.5**%p AP^{kp} and **+0.2**%p AP^{kp} without any delicately designed heuristic components. It shows that our MDPose is good at distinguishing multiple overlapping instances, which is a challenging real-world occlusion scenario.

MS COCO. Table 5 compares our MDPose with other SOTA methods on COCO test-dev set. The FPS is measured on the identical hardware if possible. Our MDPose shows the fastest inference speed with a comparable accuracy among the compared methods. Particularly, it achieves a better trade-off between the accuracy and speed compared to other single-stage instance-aware methods. Compared to FCPose, our MDPose speeds up considerably by **+8.4** FPS and **+5.0** FPS with the same backbone ResNet-50 and ResNet-101, respectively. Even with ResNet-101 which is heavier than ResNet-50, our MDPose outperforms FCPose with ResNet-50 in the inference speed by **+0.2** FPS. Compared to CMU-Pose, a representative real-time bottom-up method in multi-person pose estimation, ours achieves better accuracy and speed by a large margin. Furthermore, compared to MDN$_3$ which leverages a mixture model for multi-person pose estimation like us, our MDPose shows much improved performance in both the accuracy and inference speed, e.g. **+1.1**%p AP^{kp} and **+21.7** FPS with ResNet-50 and **+2.1**%p AP^{kp} and **+13.5** FPS with ResNet-101. Our work suggests a way for a more effective application of the mixture model in multi-person pose estimation with a much simpler architecture. The qualitative results are provided in the appendix.

5 Conclusion

Our MDPose achieves a simple pipeline by eliminating additional instance identification processes via a mixture model. The high-dimensional joint distribution of human keypoints can be learned efficiently by a simple yet effective training strategy RKG, which alleviates the underflow problem caused by the curse of dimensionality and leads to successful learning of relations between keypoints. As a result, it enables much more robust estimation under the condition of severe occlusion. Furthermore, since a mixture component corresponds to an instance, our MDPose performs instance-aware keypoint estimation without bells and whistles, enabling real-time applications. Our proposed MDPose achieves the state-of-the-art performance under the occlusion condition and is superior to other methods in the inference speed while achieving comparable accuracy. Our work shows a strong potential of a mixture model in the multi-person pose estimation and opens a way toward a much simpler pipeline for following researches.

6 Ethics statement

Our proposed MDPose enables real-time estimation of human pose with competitive accuracy, especially under the condition of a severe occlusion. Therefore, it has a great potential for a wide range of application such as a falling detection, sports teaching, customer counting and so on. Meanwhile, the effort to prevent our work from being used with malicious intention, such as illegal surveillance, should be made.

7 Reproducibility statement

We describe our training details in the Training paragraph of Section 4.1, and the code will be released in the near future.

References

Christopher M Bishop. Mixture density networks. Technical report, Citeseer, 1994.

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*, pp. 7291–7299, 2017.

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*, pp. 7103–7112, 2018.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2334–2343, 2017.

Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14676–14686, June 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *European Conference on Computer Vision*, pp. 718–734. Springer, 2020.

Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3122–3131, 2021.

Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10863–10872, 2019.

Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021.

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*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9034–9043, 2021.

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.

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

Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *European Conference on Computer Vision (ECCV)*, September 2018.

Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *European Conference on Computer Vision*, pp. 488–504. Springer, 2020.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation, 2019a. URL `https://arxiv.org/abs/1911.07451`.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019b.

Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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)*, pp. 466–481, 2018.

Nan Xue, Tianfu Wu, Gui-Song Xia, and Liangpei Zhang. Learning local-global contextual adaptation for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13065–13074, 2022.

Qi Ye and Tae-Kyun Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Jaeyoung Yoo, Hojun Lee, Inseop Chung, Geonseok Seo, and Nojun Kwak. Training multi-object detector by estimating bounding box distribution for input image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3437–3446, October 2021.

Jaeyoung Yoo, Hojun Lee, Seunghyeon Seo, Inseop Chung, and Nojun Kwak. Sparse mdod: Training end-to-end multi-object detector without bipartite matching, 2022. URL `https://arxiv.org/abs/2205.08714`.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

A Appendix

A.1 Analysis of the distribution of mixture components

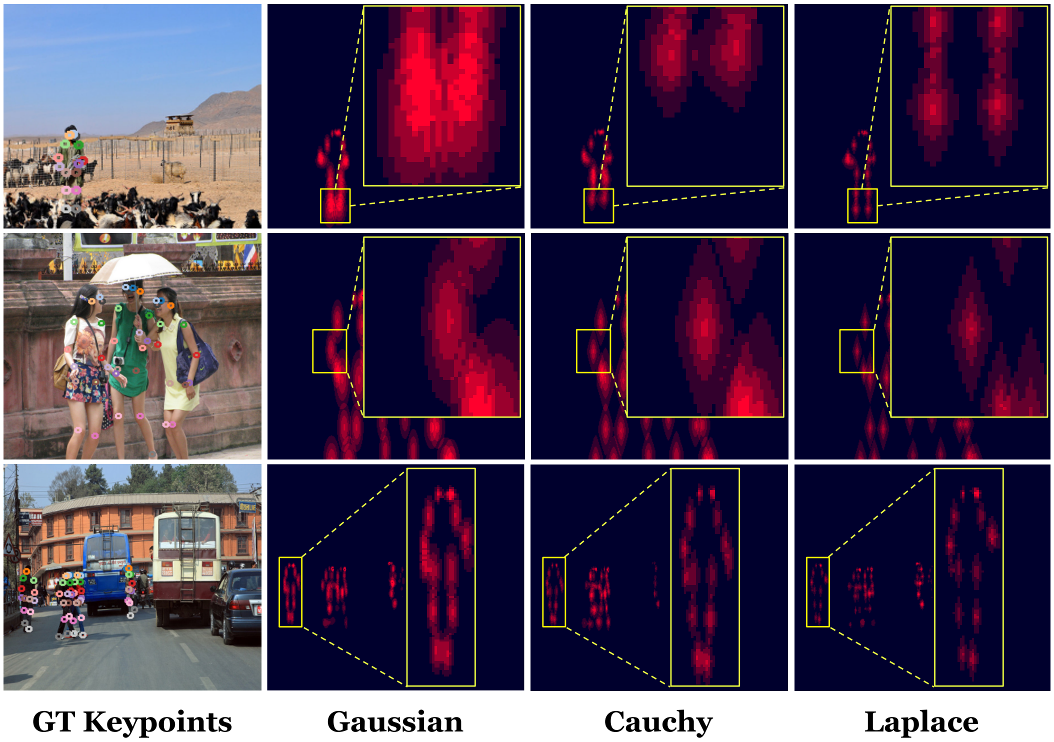


Figure 7: Visualization of estimation results with different mixture distributions of MDPose.

Table 6: Mixture model of different exponential distributions. The Laplace is more suitable than the others for multi-person pose estimation.

Dist.	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}	Underflow R.
Gaussian	50.5	79.7	54.0	41.1	63.8	0.184
Cauchy	50.6	79.6	54.1	41.4	63.5	**0.0**
Laplace	**51.5**	**80.4**	**55.1**	**42.0**	**64.7**	0.086

Table 6 shows the accuracy and underflow ratio of different mixture distributions. The MDPose with Laplace mixture distribution outperforms the one with either the Gaussian or Cauchy with a noticeable gap of AP^{kp}. Since the tails of Laplace and Cauchy fall off **less sharply** than the Gaussian, they are relatively free from the underflow problem. Furthermore, as the tails of Laplace fall off **more rapidly** than the Cauchy and it has a **sharper peak**, it leads to more efficient weighting for good and bad estimations during the training process. As demonstrated in Figure 7, the Laplace mixture distribution enables more accurate localization of human keypoints than the respective mixture distributions of the Gaussian and Cauchy.

A.2 Analysis of grouping randomness through visualization

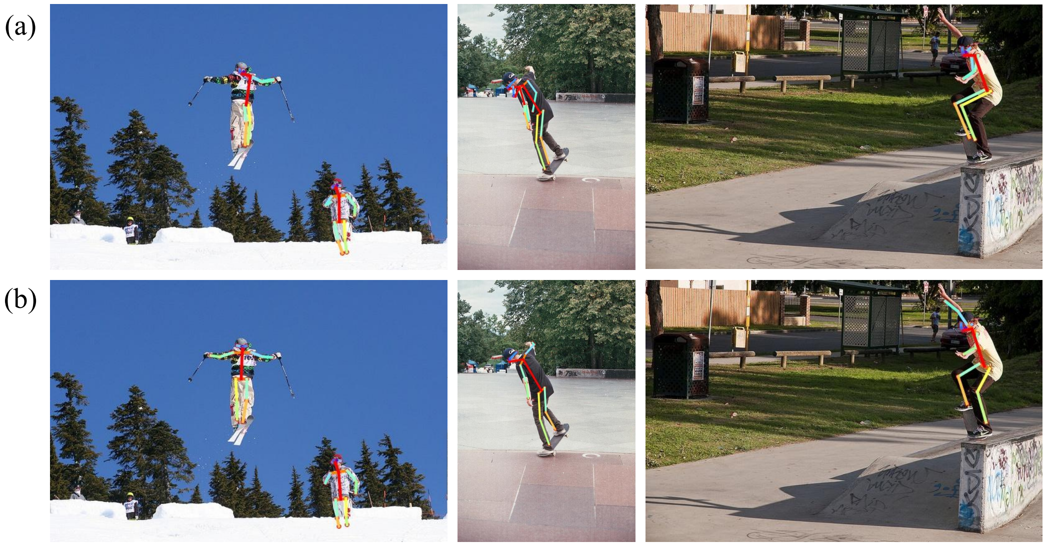


Figure 8: Visualization of our MDPose with (a) non-random grouping and (b) RKG.

As mentioned in Section 4.2 in the main paper, our proposed RKG strategy enables learning of the overall joint distributions of all keypoints while the non-random grouping learns only the joint distributions of each pre-defined keypoint groups. Figure 8 shows the qualitative results of our MDPose with (a) non-random grouping and (b) RKG. The results are obtained from the MDPose (ResNet-50) with $K_g = 3$, $N_g = 6$ and 320x320 input size on the COCO validation set. As shown in Figure 8(a), the model trained by non-random grouping has a difficulty in differentiating the left and right of limbs, due to lack of learning the overall relationship between every keypoint. On the contrary, the MDPose trained by RKG (Figure 8(b)) shows superior performance with well-distinguished left and right of limbs.

A.3 QUALITATIVE RESULTS

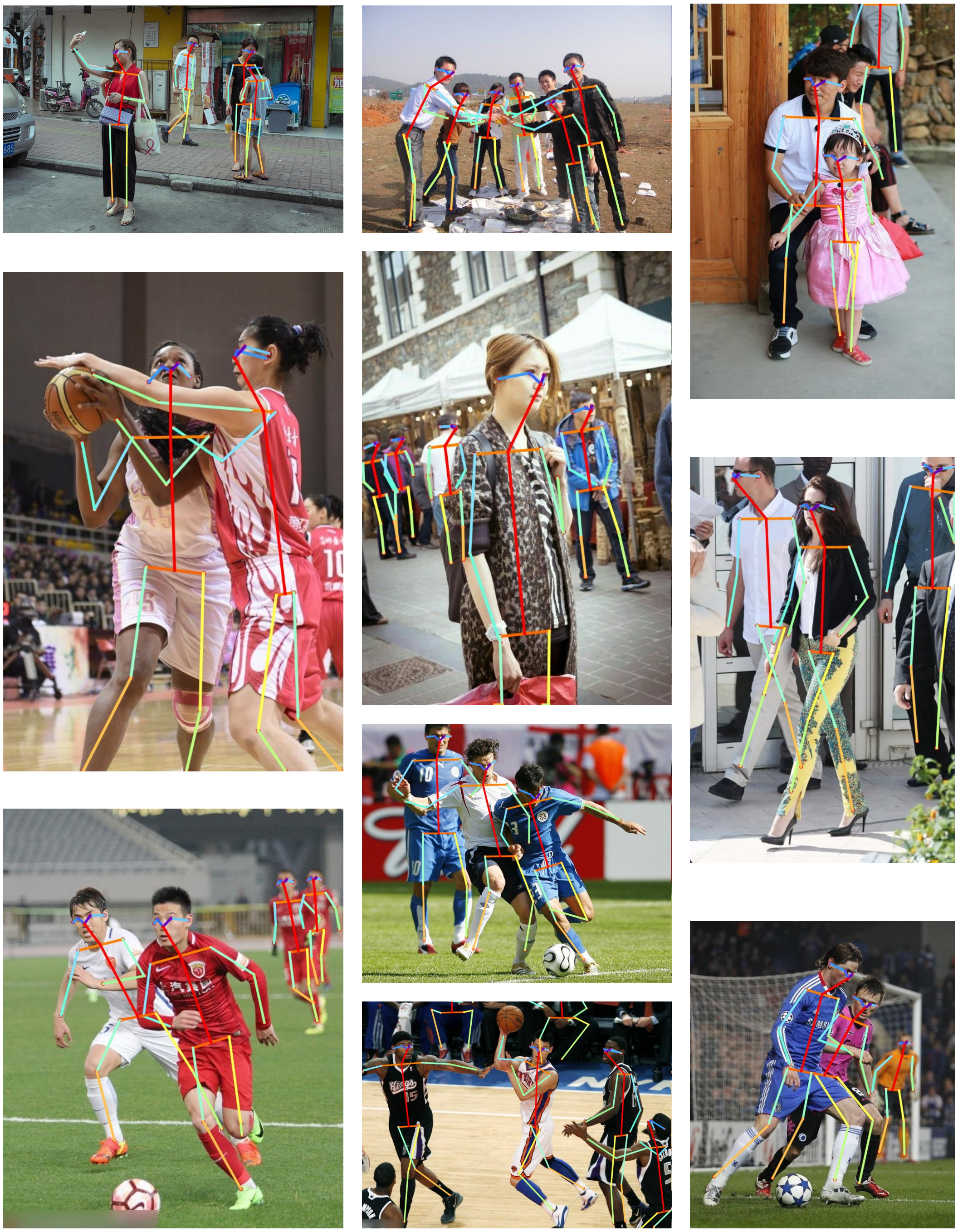


Figure 9: Qualitative results of MDPose (ResNet-101) on OCHuman validation set, with $K_g = 3$, $N_g = 6$ and 896x896 input size.

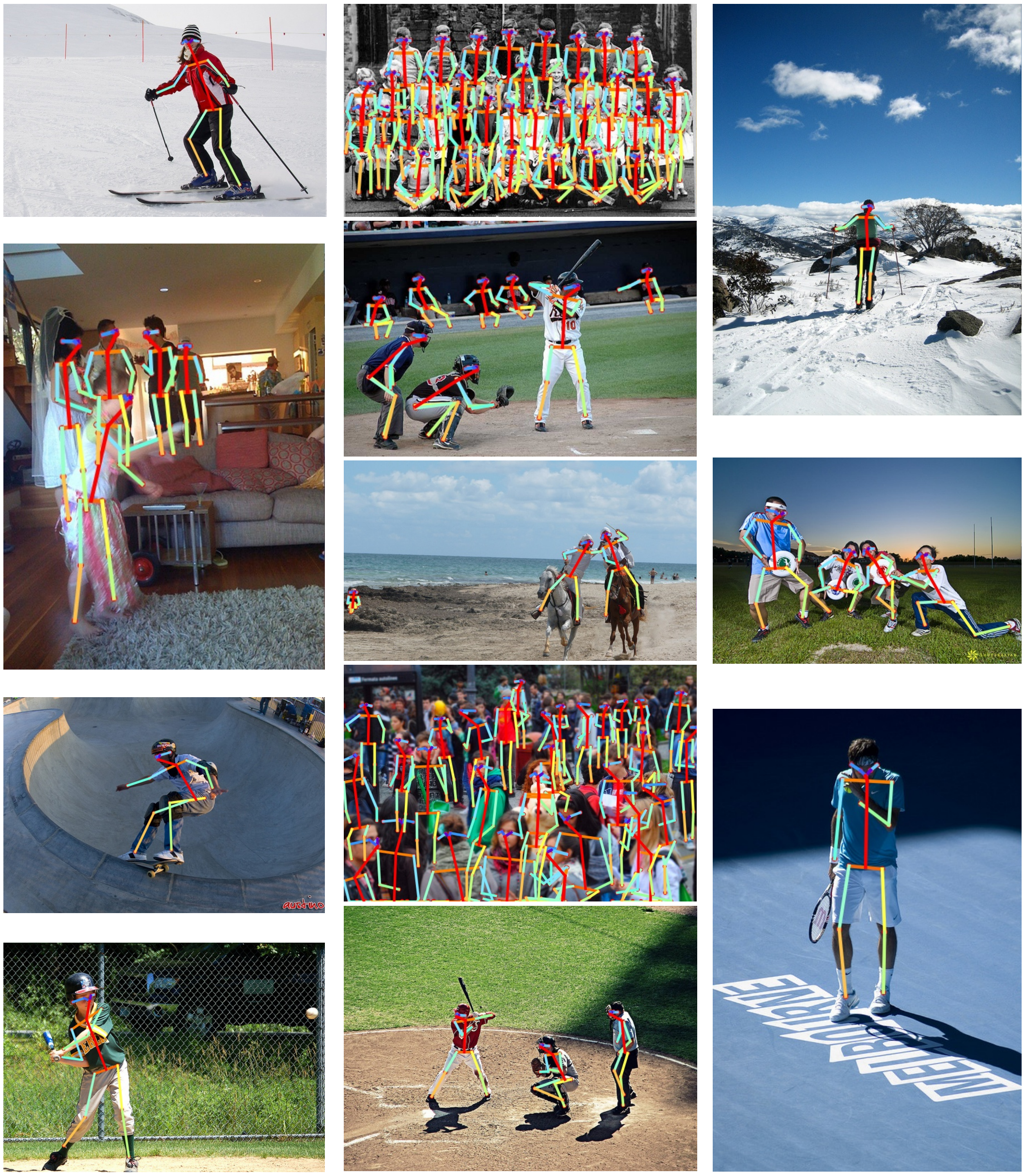


Figure 10: Qualitative results of MDPose (ResNet-50) on COCO validation set, with $K_g = 3$, $N_g = 6$ and 896x896 input size.