

Distribution-Aware Single-Stage Models for Multi-Person 3D Pose Estimation

Zitian Wang^{1,3}

Xuecheng Nie³

Xiaochao Qu³

Yunpeng Chen³

Si Liu^{1,2,†}

¹Institute of Artificial Intelligence, Beihang University, China

²Hangzhou Innovation Institute, Beihang University, China

³MT Lab, Meitu Inc., China

wangzt.kghl@gmail.com {nxc, qxc, cyp5}@meitu.com liusi@buaa.edu.cn

Abstract

In this paper, we present a novel Distribution-Aware Single-stage (DAS) model for tackling the challenging multi-person 3D pose estimation problem. Different from existing top-down and bottom-up methods, the proposed DAS model simultaneously localizes person positions and their corresponding body joints in the 3D camera space in a one-pass manner. This leads to a simplified pipeline with enhanced efficiency. In addition, DAS learns the true distribution of body joints for the regression of their positions, rather than making a simple Laplacian or Gaussian assumption as previous works. This provides valuable priors for model prediction and thus boosts the regression-based scheme to achieve competitive performance with volumetric-base ones. Moreover, DAS exploits a recursive update strategy for progressively approaching to regression target, alleviating the optimization difficulty and further lifting the regression performance. DAS is implemented with a fully Convolutional Neural Network and end-to-end learnable. Comprehensive experiments on benchmarks CMU Panoptic and MuPoTS-3D demonstrate the superior efficiency of the proposed DAS model, specifically 1.5x speedup over previous best model, and its stat-of-the-art accuracy for multi-person 3D pose estimation.

1. Introduction

Estimating 3D poses of multiple persons from a single RGB image is a fundamental yet challenging task in computer vision, which targets at localizing 3D positions of persons and their body joints in the camera space. Recently, it has drawn much attention thanks to the broad applications in AR/VR [2, 14], Gaming [30, 31, 33], Human-Computer Interaction [7, 32], etc.

Prior works address this task via two-stage strategies. They either adopt the *top-down* scheme [22, 28, 35, 41] that

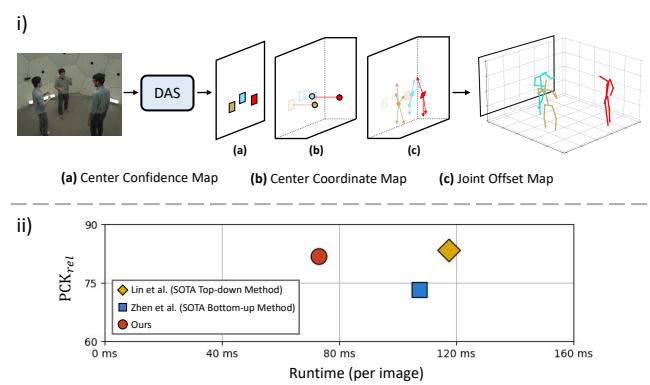


Figure 1. Overview of our Distribution-Aware Single-stage (DAS) model for multi-person 3D pose estimation. i) The brief pipeline of DAS. ii) Comparison with state-of-the-arts on MuPoTs-3D. DAS outperforms two-stage models in efficiency. It achieves superior performance with SOTA bottom-up method while competitive with SOTA top-down one. For PCK_{rel}, the higher, the better.

first localizes the absolute 3D person positions and then separately estimates the root-relative body joints for each person; or the *bottom-up* scheme [8, 21, 47] that in the first stage detects all 3D body joints and groups them into the corresponding persons in the second stage. Although achieving good accuracy, these methods suffer redundant computation and complex postprocessing, caused by the sequential management for person position and body joint localization in a two-stage manner. This leads to unsatisfied efficiency for deployment in real scenarios.

Given the above motivation, we propose to simplify the pipeline for multi-person 3D pose estimation, further pushing forward the frontier of its applications. Inspired by schemes in the 2D counterpart [23, 40, 48], we aim to design a single-stage model to simultaneously localize 3D person positions and body joints. However, the extension from 2D to 3D scenes is non-trivial, due to the ill-posed setup for deriving depth information from a monocular RGB image without prior knowledge on the data distribution.

To achieve the goal, we present a novel Distribution-Aware Single-stage (DAS) model in this paper. The pro-

[†]Corresponding author.

posed DAS tackles the ill-posed problem of multi-person 3D pose estimation from two aspects: 1) DAS represents the 3D pose with a 2.5D human center together with 3D center-relative joint offsets. This adapts to direct depth prediction from the image domain as well as enables to unify the 3D localization of person position and body joints, making the monocular-based one-pass solution feasible; 2) DAS learns the true distribution of body joints during the model optimization. This provides valuable guidance to predict the joint position, thus boosting the performance of the regression-based scheme. To alleviate the distribution estimation difficulty, DAS exploits a recursive update strategy to progressively approach to the targets. In this way, DAS can efficiently generate accurate 3D poses of multiple persons from a single RGB image.

In particular, we build DAS with a regression-based pipeline that outputs the 3D human poses via a single forward inference from an input image. As depicted in Figure 1 i) (a) and (b), DAS models the human center with a center confidence map and a center coordinate map. DAS uses the center confidence map for localizing the projected human centers in 2D image coordinate space, while the center coordinate map for pixel-wisely estimating the absolute center positions in 3D camera coordinate space. DAS utilizes joint offset maps to densely encode 3D center-relative locations of body joints, as shown in Figure 1 i) (c). DAS can produce these three kinds of maps in parallel and then easily reconstruct multiple 3D human poses with them, avoiding redundant computation and complex association. With this compact single-stage pipeline, DAS can achieve superior efficiency over prior two-stage ones.

To optimize the regression-based model, prior works always adopt the conventional L1 or L2 loss. However, researches [12] have proven that this kind of supervision actually makes a simple Laplacian or Gaussian assumption on the data distribution, which is far away from the true one. In contrast, DAS learns the underlying distribution of 3D body joint locations via exploiting the normalizing flows [1, 5, 27]. This helps to derive a suitable distribution for model output, thus providing valuable priors to guide the regression of body joint coordinates. Together with the pose regression modules, DAS optimizes the distribution learning module through the maximum likelihood estimation in the training phase, while removes it in the inference phase. In this way, DAS boosts the regression performance without additional computation cost. Besides, DAS iteratively updates joint offsets by ensemble of informative predictions around regression targets to further facilitate the localization of body joints. With this distribution-aware design, DAS can achieve outperforming accuracy with bottom-up methods while competitive with top-down ones, as shown in Figure 1 ii).

We implement DAS with a fully Convolutional Neu-

ral Network, which is end-to-end learnable. Comprehensive experiments on benchmarks CMU Panoptic [10] and MuPoTS-3D [21] show the superiority of our DAS model. In summary, our contribution are in three folds: (1) We present a novel single-stage model for estimating 3D poses of multiple persons from a monocular RGB image, which overcomes the drawbacks on computation cost and model complexity occurred in two-stage methods; (2) We present a novel distribution-aware way to boost the body joint regression in a recursive manner; (3) We set new state-of-the-art on multiple benchmarks with superior efficiency.

2. Related Works

2.1. Two-stage 3D Human Pose Estimation

Approaches for monocular multi-person 3D pose estimation can be divided into two categories, the top-down methods and bottom-up methods. The top-down methods require additional human detector to provide human position prior for downstream single person pose estimator to generate individual 3D poses [24, 34, 35, 41]. [28, 29] introduce a pose proposals network to generate human bounding boxes and a series of human pose hypotheses, then the pose hypotheses are refined based on the cropped ROI features. To estimate the camera-centric human pose, [22] disentangles absolute human depth estimation and human-centric pose estimation by using separate models. [37] propose hierarchical ordinal relations objective along with virtual view sample strategy to constrain the instance level and joint level depth values. [15] utilize GNN to infer the human depth based on the human-centric joint features. Although achieving high accuracy, top-down methods suffers high computation cost as person number increases.

Another group of approaches investigate multi-person 3D pose estimation in a bottom-up manner. Similar to the case in 2D pose estimation, the bottom-up methods are commonly composed of joint localization and joint association. [21] combines the joint location maps and the occlusion-robust pose-maps to infer the 3D poses based on the results of [4]. [8] use volumetric heatmaps to model joint locations with an encoder-decoder network for feature compression, and a distance-based heuristic is applied to retrieve 3D pose for each person. [45] propose to group human joints according to body part scores based on integrated 2D and 3D information. [47] develop 3D part affinity field for depth-aware part association, and utilizes a refine network to refine the 3D pose given predicted 2D and 3D joint coordinates. Though avoiding repeated single person pose estimation, these methods requires a second association stage for joint grouping.

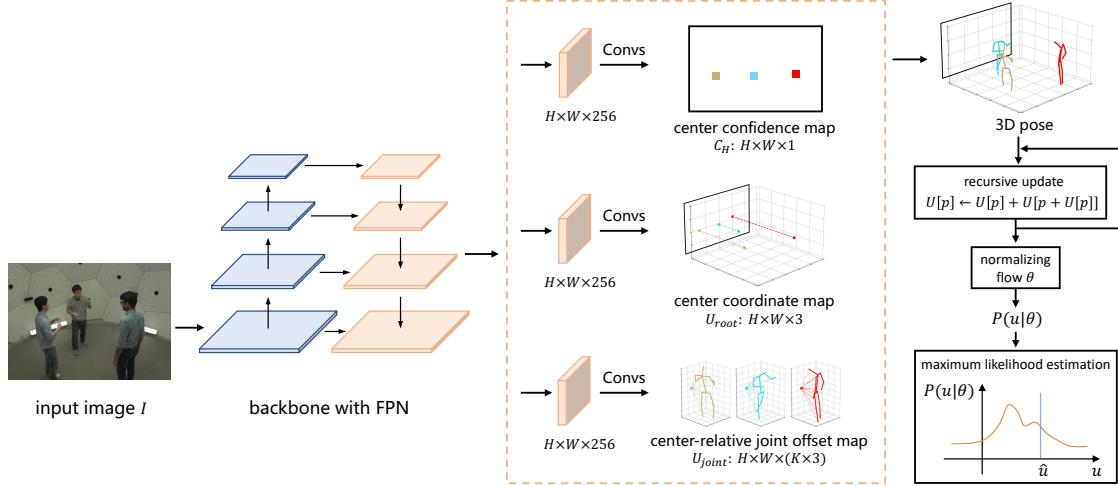


Figure 2. Illustration of the framework of DAS. The input image is fed to 2D CNN backbone for feature extraction. A followed FPN neck produces multi-level feature maps of different sizes. Shared prediction heads upon the multi-level FPN features are responsible for human center detection, center coordinate regression and center-relative joint offset regression respectively. The 3D joint locations are reconstructed based on obtained center coordinates and joints offsets. The probability distribution of 3D joint locations is modeled by normalizing flow with recursive update. Maximum likelihood estimation is applied to assist the learning process.

2.2. Single-stage 2D Human Pose Estimation

There exists some researches that employ a single-stage pipeline for 2D multi-person human pose estimation. The single-stage methods treat the pose estimation as parallel human center localizing and center-to-joint regression problem [23, 40, 48]. Instead of separating joints localizing and grouping, these methods predict each of the joint offsets from the detected center points, which is usually set as the root joint of human. Since the joint offsets are directly correlated to estimated center points, this formulation avoid the manually designed grouping post-processing and is end-to-end trainable. Zhou et al. [48] direct regress the joint locations from human center. Nie et al. [23] introduce a hierarchical structured pose representation to assist the long-range regression. Wei et al. [40] propose to regress the joint locations from point-set anchors which serve as prior of basic human poses. These single-stage methods have high efficiency for deployment.

2.3. Normalizing Flows for Human Pose Estimation

Some recent works introduce normalizing flow for 3D human pose and shape estimation. Xu et al. [42] propose a statistical 3D articulated human shape model with normalizing flow representations for skeleton kinematics. Zanfir et al. [43] introduce the kinematic latent normalizing flow representation for weakly supervised 3D human pose and shape reconstruction. Biggs et al. [3] and Wehrbein et al. [39] propose to model the ambiguities and occlusions in 3D human pose and shape estimation by normalizing flow. Kolotouros et al. [11] learn a mapping from 2D image to a

distribution of 3D body meshes based on conditional normalizing flow. Different from the above methods, Li et al. [12] use normalizing flow to model the underlying distribution of joint location, and propose to optimize the parameters by residual log-likelihood estimation.

Inspired by [12], we propose a recursive flow based optimization scheme for 3D joint offset regression. The distribution of joint location is modeled with normalizing flow and updated in a recursive manner. Then the regression model is optimized by maximum likelihood estimation.

3. Method

3.1. Overview

In this section, we will give an overview on the proposed Distribution-Aware Single-stage (DAS) model for multi-person 3D pose estimation.

Given an RGB image I as input, DAS estimates the joint locations $H^{img} = \{J_k^{img} = (x_k^{img}, y_k^{img}, d_k) | k \in [1 \dots K]\}$ in image coordinate system for every person. These locations are composed of the 2D image coordinates and camera-centric depth values. DAS factorizes the multi-person 3D pose estimation as simultaneously human center localization and center-relative joint offset regression. The illustration of the framework is shown in Figure 2. For an input image, the CNN backbone with FPN [16] first extracts multi-level feature maps $\{F_1, \dots, F_L\}$ to deal with humans of different scales. Three parallel branches on the top of the feature maps serve for human center detection, center coordinate regression and center-relative joint offset regression respectively. The multi-person 3D poses can be reconstructed

through combining the outputs of these three branches. To improve the joint localization ability in this regression based framework, a recursive flow based optimization scheme is proposed for accurately modeling the probability distribution of joint locations. Details are given in sections below.

3.2. Human Center Localization

Ground truth assignment Given one person with joint locations H^{img} , it is assigned to FPN feature map F_l with downsample stride s_l according to the max joint distance from root $r_{max} = \max_k(\sqrt{(x_k^{img} - x_{root}^{img})^2 + (y_k^{img} - y_{root}^{img})^2})$. Specifically, we set the regression range for F_l as $[m_{l-1}, m_l]$. If r_{max} is within the range of $[m_{l-1}, m_l]$, H^{img} will be assigned to the l th feature map. Then the coordinates would be scaled to $H = \{j_k = (x_k, y_k, d_k) | k \in [1...K]\}$ to match the downsample stride, s.t., $x_k = x_k^{img}/s_l$ and $y_k = y_k^{img}/s_l$.

Human center detection We consider the root joint (i.e., pelvis) j_{root} as the human center. The center confidence map C_H with shape $[H \times W \times 1]$ measures whether the pixels represent any human center. The human center detection is treated as binary classification. The nearest N_{pos} pixels around each human center are set as positive samples $\{p_0, \dots, p_{N_{pos}}\}$. The ground truth confidence is 1 for positive samples and 0 for the others. The loss function is:

$$L_{cls} = \text{FocalLoss}(C_H, \hat{C}_H), \quad (1)$$

where C_H is the predicted center confidence map, \hat{C}_H is the ground truth and FocalLoss is proposed in [17]. Besides, a centerness branch is added to assess the center quality. Design of the centerness target and the loss function $L_{centerness}$ follows [36].

Center coordinate regression Given the positive sample $p = (x_p, y_p)$ w.r.t the center of human H , the center coordinates are regressed from p . With the predicted center coordinate map U_{root} shaped of $[H \times W \times 3]$, the regression target for the center $j_{root} = (x_{root}, y_{root}, d_{root})$ at p is set as the offset $U_{root}[p] = (x_{root} - x_p, y_{root} - y_p, d_{root})$. L1 loss is adopted for this regression task:

$$L_{root} = \sum_p ||U_{root}[p] - \hat{U}_{root}[p]||_1, \quad (2)$$

where $U_{root}[p]$ is the predicted center coordinate offset at p and $\hat{U}_{root}[p]$ is the ground truth.

3.3. 3D Body Joint Localization

Center-relative joint offset regression Unlike heatmap based formulation that needs a second association stage for joint identities, this regression based formulation makes joint localization and identification a holistic process. The

center-relative joint offsets are regressed from each positive sample p . We predict center-relative joint offset map $U_{joint} = \{U_1, \dots, U_K\}$, where U_k shaped of $[H \times W \times 3]$ is responsible for the 3D center-relative offset of the k th joints. The regression target for $H = \{j_k | k \in [1...K]\}$ at p is set as $U_k[p] = j_k - j_{root}$. Traditional regression loss (e.g., L1 loss) can be adopted for optimization. The objective can be written as:

$$L_{pose} = \sum_k \sum_p ||U_k[p] - \hat{U}_k[p]||_1, \quad (3)$$

where $U_k[p]$ is the predicted center-relative joint offset at p and $\hat{U}_k[p]$ is the ground truth.

Joint location distribution modeling Compared to the heatmap based formulation which commonly represents the joint location as Gaussian distribution, the regression based formulation represents the joint location in a deterministic form. Denote the center-relative joint location as u , the deterministic representation could be more inclined to be affected by label noises, occlusions and invisibility. Given this consideration, we employ normalizing flow [1, 5, 6, 27] to represent the center-relative joint location in a probability distribution form: $u \sim P(u)$.

Following [12], we model the distribution $P(u)$ with reparameterization. Specifically, $P(u)$ can be obtained by scaling and shifting z from a zero-mean distribution $z \sim P_Z(z)$ with transformation function $u = \bar{u} + \sigma \cdot z$, where \bar{u} stands for the expectation of joint location and σ indicates the scale of the distribution. Given this transformation function, the density function of $P(u)$ can be calculated as:

$$\log P(u) = \log P_Z(z) - \log \sigma. \quad (4)$$

Therefore, rather than regressing the deterministic center-relative joint location u , we regress the expectation \bar{u} and the scale indicator σ instead. While $P_Z(z)$ can be modeled by a normalizing flow model (e.g., real NVP [6]). Note that only the expectation \bar{u} is needed to calculate the eventual joint location in the inference phase.

Recursive flow based optimization Since the feature used to infer the human pose is picked from the human center position, it is less representative for joints that are far away from the center. This spatial nonalignment between feature and target can cause larger errors than the counterparts in bounding box regression [17, 26, 36] due to the complexity of human body structure.

To alleviate this problem, we propose a recursive update strategy to iteratively optimize the location expectation \bar{u} . The idea is depicted in Figure 3. Take the k th joint offset map U_k as example, given the initial prediction $\bar{u} = U_k^n[p]$ from positive sample p , it is updated by the *local prediction*

from $p + U_p^k[p]$:

$$U_k^{n+1}[p] \leftarrow U_k^n[p] + \mathbf{B}(U_k^{local}[p + U_k^n[p]]), \quad (5)$$

where $\mathbf{B}(\cdot)$ indicates bilinear interpolation function to obtain values from non-integer coordinates.

For *local prediction*, we can use the same prediction map as U_k^n , i.e., $U_k^{local} = U_k^n$. This formulation allows pixels around target joint to produce high quality joint offsets. By this means, \bar{u} is updated in a recursive form as in Figure 3c:

$$\begin{aligned} U_k^{n+1}[p] &\leftarrow U_k^n[p] + U_k^n[p + U_k^n[p]] \\ \bar{u} &\leftarrow U_k^{n+1}[p], \end{aligned} \quad (6)$$

where the interpolation operation is omitted for simplicity.

Moreover, for better modeling \bar{u} , we consider another multi-source update strategy to approximate the expectation. In this case, $U_k^{local}[t]$ is calculated by ensembling multiple predictions sampled around t as in Figure 3d:

$$\begin{aligned} U_k^{local}[t] &= \mathbf{E}_{d \sim P_D(d)}(d + U_k[t + d]) \\ &\approx \sum_m P_D(d_m)(d_m + U_k[t + d_m]), \end{aligned} \quad (7)$$

where d_m and $P_D(d_m)$ are the sample location and probability for the m th sample source generated by MLP.

The recursive update strategy can be implemented with convolutional layers and interpolation layers. By stacking update layers on the top of joint offset map, \bar{u} can be optimized progressively without modification on model pipeline. Besides, this recursive form avoids manually target assignment at each position in the joint offset map.

To take the advantage of the distribution-aware representation, we exploit maximum likelihood estimation for parameter optimization in the training phase. After obtaining \bar{u} and σ , joint location distribution can be expressed as in Equation (4). If $P_Z(z)$ is modeled by a normalizing flow model θ , the maximum likelihood estimation objective can be written as:

$$\begin{aligned} L_{mle} &= -\log P(u)|_{u=\hat{u}} \\ &= -\log P_Z(\hat{z}|\theta) + \log \sigma, \end{aligned} \quad (8)$$

where $\hat{z} = (\hat{u} - \bar{u})/\sigma$, \bar{u} is the estimated expectation of center-relative joint location, and \hat{u} is the ground truth location. Through L_{mle} , the distribution $P_Z(z|\theta)$ can be learned together with \bar{u} .

In this work, we further follow [12] to use residual log-likelihood estimation (RLE). RLE factorizes the distribution $P_Z(z)$ into one prior distribution $Q_Z(z)$ (e.g., Laplace distribution and Gaussian distribution) and one learned distribution $G_Z(z|\theta)$. The RLE objective can be written as:

$$L_{rle} = -\log Q_Z(\hat{z}) - \log G_Z(\hat{z}|\theta) + \log \sigma. \quad (9)$$

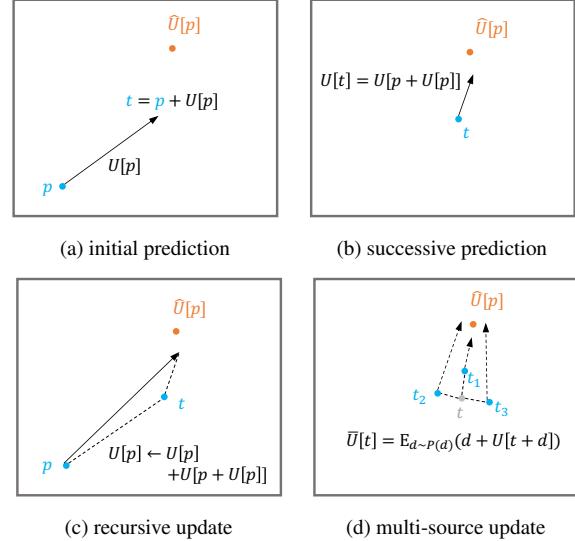


Figure 3. Illustration of the recursive update adopted in the proposed DAS model. Subscripts are omitted for simplicity.

We recommend the reader to refer to the original paper for more details. In experiments, we implement RLE objective to replace L1 loss for L_{pose} .

Reconstruction of 3D pose The joint coordinates can be obtained easily by adding up the coordinates of human center and the center-relative joint offsets without other association step. Thus the 3D locations of human joints can be inferred in a single forward pass.

3.4. Training and Inference

Training During training, we first transform the 3D joint coordinates into image coordinate system using the camera intrinsic parameters. Since the same object can have different depth values in images captured from different cameras, it is hard for model to learn an absolute representation of object depth directly. Therefore we follow the previous work [22] to use normalized depth $d_{norm} = d/f$ as target depth for center coordinate regression, where f is the camera focal length. Besides, for center-relative pose offset, we do not normalize the depth value for training stabilization. The overall objective is:

$$L = L_{cls} + \lambda_1 L_{centerness} + \lambda_2 L_{root} + \lambda_3 L_{pose}, \quad (10)$$

where $\lambda_1, \lambda_2, \lambda_3$ are loss weights.

Inference The input image is first fed into the model to produce all the intermediate results. Then the positions where the center confidences are higher than a threshold (set as 0.05) are chosen as positive samples. Corresponding center coordinates and center-relative joint offsets are taken from the positive samples to form the camera-centric

center type	recursive update	MLE	larger backbone	MPJPE			MPJPE
bbox				65.3			62.5
root				62.5			58.6
root	✓			57.6	✓	1	58.2
root		✓		58.8	✓	✓	57.9
root	✓	✓		56.3	✓	✓	57.6
root	✓	✓	✓	54.4	✓	✓	57.6

(a) Component analysis of our single-stage method.

(b) Comparison of different settings for recursive update.

Table 1. Ablation study on CMU Panoptic dataset.

method	Haggling	Mafia	Ultimatum	Pizza	Mean ↓
<i>Top-down methods</i>					
Popa et al. [25]	217.9	187.3	193.6	221.3	203.4
Zanfir et al. [44]	140.0	165.9	150.7	156.0	153.4
Wang et al. [37]	50.9	50.5	50.7	68.2	55.1*
<i>Bottom-up methods</i>					
Zanfir et al. [45]	72.4	78.8	66.8	94.3	78.1*
Fabbri et al. [8]	45	95	58	79	69
Zhen et al. [47]	63.1	60.3	56.6	67.1	61.8
Ours	53.3	51.2	49.1	61.5	53.8

Table 2. Comparison with SOTAs on CMU Panoptic dataset in MPJPE. The top rows are two-stage methods and the bottom row is our single-stage methods. * means that Mean MPJPE is recalculated by averaging over activities following the standard practice in [44].

3D poses. Non-maximum suppression is adopted to reduce redundant pose hypotheses.

4. Experiments

4.1. Datasets

CMU Panoptic CMU Panoptic is a large-scale real-word indoor 3D human pose dataset containing 65 video sequences of daily activities [10]. We follow the evaluation protocol proposed by [44] to use videos from HD cameras of indices 16 and 30. For evaluation, 9600 frames from 4 activities (*Haggling*, *Mafia*, *Ultimatum* and *Pizza*) are chosen as test set. For training, we use a mixed dataset composed of COCO [18] and video sequences from 3 activities (*Haggling*, *Mafia* and *Ultimatum*). The training set has no overlap with the test set. For images with only 2D pose annotations, the depth information is ignored in loss calculation. We follow the previous works to use Mean Per Joint Position Error (MPJPE) for performance evaluation [38, 44]. MPJPE is calculated after aligning the pose by the root joint.

MuCo-3DHP and MuPoTS-3D MuCo-3DHP and MuPoTS-3D are multi-person 3D pose estimation datasets for training and evalution respectively proposed in [21]. The training set MuCo-3DHP is a large-scale synthesized dataset. The images are generated by randomly compositing the persons from single person 3D pose estimation dataset MPI-INF-3DHP [19]. The test set MuPoTS-3D is a realistic dataset captured from real-world outdoor

scenes which is annotated with marker-less motion capture system. MuPoTS-3D contains 20 video sequences. Each video has up to 3 subjects. We use the mixed datasets composed of COCO and MuCo-3DHP for training, and use MuPoTS-3D for evaluation. We follow the previous works to use the 3D Percentage of Correct Keypoints (3DPCK) for performance evaluation [22, 46]. In particular, PCK_{rel} is used to evaluate the predictions after root alignment, and PCK_{abs} is used to evaluate the predictions under camera coordinate system. One joint is treated as correct if it is within 15cm distance from the matched ground truth.

4.2. Ablation Study

We conduct ablation study on CMU Panoptic dataset. The model is implemented with ResNet-50 [9] and FPN [16]. The experimental results are shown in Table 1.

We first analyse different components in our method, as shown in Table 1a. Compared to the bounding box center, using root joint as human center is more beneficial for body joint regression, which brings improvement of 2.8mm in MPJPE. When equipped with recursive update strategy, 4.9mm improvement in MPJPE is obtained. If the joint location distribution is modeled by normalizing flow, maximum likelihood estimation (MLE) during training effectively boosts the localization ability and improves MPJPE by 3.7mm. When combining the recursive update with MLE, MPJPE achieves 56.3mm, which exceeds baseline by 6.2mm. With a larger backbone [13], the proposed method can achieve a lower MPJPE of 54.4mm.

method	PCK _{rel} ↑	PCK _{abs} ↑
<i>Top-down methods</i>		
Rogez et al. [29]	70.6	n/a
Moon et al. [22]	81.8	31.5
Wang et al. [37]	82.0	43.8
Lin et al. [15]	83.7*	35.2*
<i>Bottom-up methods</i>		
Mehta et al. [21]	65.0	n/a
Mehta et al. [20]	70.4	n/a
Zhen et al. [47]	73.5	35.4
Ours	82.7	39.2

Table 3. Comparison with SOTAs on MuPoTS-3D dataset. PCK_{rel} and PCK_{abs} are reported for *all groundtruths*. * means the value is reported on *matched groundtruths*.

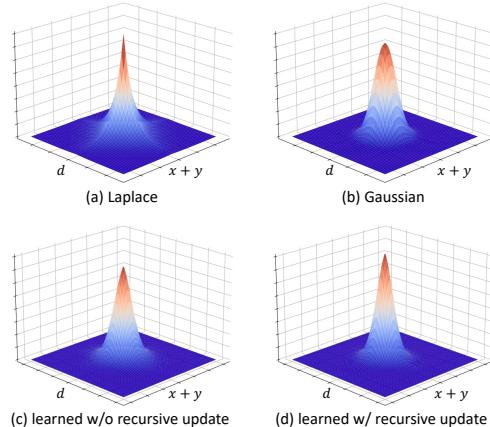


Figure 4. Visualization of the learned distribution.

The detailed comparison of different settings for recursive update is shown in Table 1b. By applying recursive update, MPJPE can be improved by 3.8mm compared to the baseline. With multi-source update, MPJPE is further improved by 0.4mm. Slight performance gains are achieved if stacking more update layers.

4.3. Quantitative Results

Comparison on CMU Panoptic We conduct experiments to compare with both the state-of-the-art top-down and bottom-up methods on CMU Panoptic. The quantitative results are reported in the Table 2. MPJPE is adopted to evaluate the performance of 3D pose estimation after root alignment. As shown in results, the proposed single-stage method outperforms previous volumetric based bottom-up methods in mean MPJPE, and achieves slightly better performance compared to previous top-down methods.

Comparison on MuPoTS-3D We evaluate the proposed method on MuPoTS-3D test set using the evaluation metric PCK_{rel} and PCK_{abs}. We compare our method with state-of-

method	runtime	MPJPE	PCK _{rel}
<i>Top-down methods</i>			
Moon et al. [22]*	107 ms	n/a	81.8
Lin et al. [15]*	118 ms	n/a	83.7
<i>Bottom-up methods</i>			
Fabbri et al. [8]	125 ms	69	n/a
Zhen et al. [47]*	108 ms	61.8	73.5
Ours	75 ms	54.6	81.0

Table 4. Runtime comparison with SOTAs. * means the speed is reproduced based on official repository.

the-art top-down and bottom-up methods in Table 3. From the experimental results, our method outperforms previous bottom-up methods especially in PCK_{rel}. This result demonstrates the effectiveness of our single-stage method in estimating the human body structure. Compared to the top-down methods, our method obtains comparable performance without a second stage for single person pose estimation and depth estimation.

4.4. Running Time Analysis

We compare the running time of our method with top-down and bottom-up methods in Table 4. In this experiment, our method is implemented with 2-stage MSPN [13], and the running time is measured on a single V100 GPU with batch size 1. Under this setting, we also give the corresponding MPJPE and PCK_{rel} on CMU Panoptic and MuPoTS-3D respectively. Compared to the top-down methods, our method achieves comparable performance with $\sim 1/3$ less runtime. Considering there are few people (less than 3 per image on average) on MuPoTS-3D, the time consumption of the top-down methods can increase rapidly in complex scenes. Compared to the bottom-up methods, our method maintains faster inference speed since there is no need for high-resolution 2D or 3D heatmaps. This property makes DAS a computational-friendly method.

4.5. Qualitative Results

We visualize the distribution learned by normalizing flow in Figure 4. The difference between learned distribution and prior distributions demonstrates the importance of distribution-aware optimization scheme. Besides, the learned distribution with recursive update has less variance than the case without, indicating the recursive update strategy assists to accurately model the underlying distribution. Also, we visualize the learned joint offset maps by 2D projection in Figure 5. As can be seen, with recursive flow based optimization, the pixels learn to regress the joint location either around human center or around target joint at prediction maps. The learning of the location distribution of long-distance joints can benefit from the more precise local predictions. In Figure 6, we provide visualization results



Figure 5. Visualization of the joint offsets by 2D projection. The top row is taken from the prediction map with downsample stride 16 and the bottom row is taken from the prediction map with downsample stride 32.

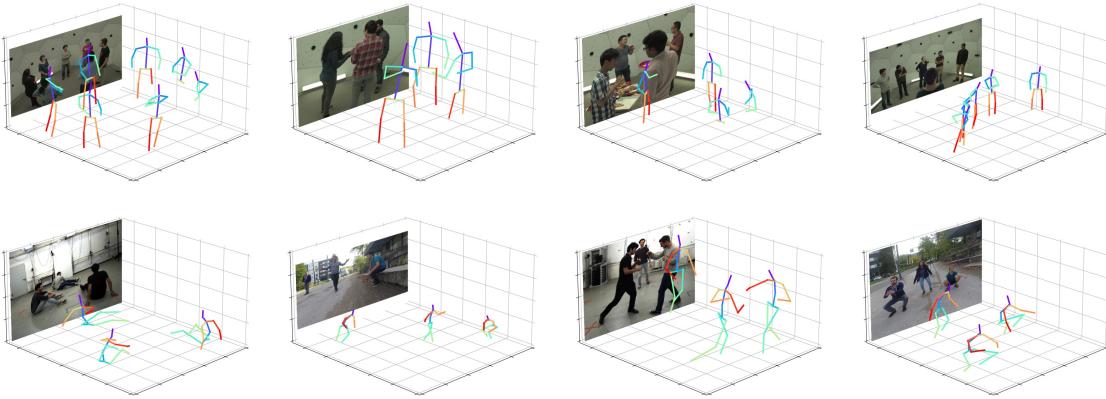


Figure 6. Qualitative results of the proposed DAS model for multi-person 3D pose estimation. Best viewed in color.

of our DAS model. We can see DAS can produce accurate predictions in various scenarios, e.g., pose changes, person truncation, clutter backgrounds.

5. Conclusions

In this paper, we present a novel Distribution-Aware Single-stage (DAS) model for solving the challenging multi-person 3D pose estimation problem. Different from existing two-stage methods, *i.e.*, top-down or bottom-up ones, the proposed DAS model can simultaneously locate the positions of persons and their body joints in camera coordinate space via a one-pass fashion. This helps to simplify the pipeline and overcome drawbacks of prior works on high computation cost and model complexity. In addition, DAS successfully introduces the normalizing flows into multi-person 3D pose estimation task and learns the joint location distribution together in the training phase. Moreover, DAS adopts a recursive flow based optimization scheme for progressively refining the location distribution. By this way, DAS derives the true underlying distribution,

thus boosting the regression performance. We conduct extensive experiments on multiple benchmarks and verify the effectiveness and efficiency of the proposed DAS model for multi-person 3D pose estimation.

Limitations The proposed DAS model is limited for estimating poses of extremely overlapped persons in an image. Since DAS represents the 3D human pose by human center and corresponding joint offsets, the overall performance is highly affected by the center detection process. When persons are extremely overlapped, their centers can also be overlapped in the image space. In this case, some occluded centers may be erroneously suppressed.

Acknowledgments

This research is partly supported by National Natural Science Foundation of China (Grant 62122010, 61876177), Fundamental Research Funds for the Central Universities, and Key R&D Program of Zhejiang Province (2022C01082).

References

- [1] JP Agnelli, M Cadeiras, Esteban G Tabak, Cristina Vilma Turner, and Eric Vanden-Eijnden. Clustering and classification through normalizing flows in feature space. *Multiscale Modeling & Simulation*, 2010. 2, 4
- [2] Hayet Belghit, Abdelkader Bellarbi, Nadia Zenati, and Samir Otmane. Vision-based pose estimation for augmented reality: a comparison study. *arXiv preprint arXiv:1806.09316*, 2018. 1
- [3] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *NIPS*, 2020. 3
- [4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2
- [5] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2, 4
- [6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 4
- [7] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 2007. 1
- [8] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *CVPR*, 2020. 1, 2, 6, 7
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [10] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2, 6
- [11] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 3
- [12] 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. 2, 3, 4, 5
- [13] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 6, 7
- [14] Huei-Yung Lin and Ting-Wen Chen. Augmented reality with human body interaction based on monocular 3d pose estimation. In *ACIVS*, 2010. 1
- [15] Jiahao Lin and Gim Hee Lee. Hdnet: Human depth estimation for multi-person camera-space localization. In *ECCV*, 2020. 2, 7
- [16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3, 6
- [17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4
- [18] 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 *ECCV*, 2014. 6
- [19] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 6
- [20] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *TOG*, 2020. 7
- [21] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 1, 2, 6, 7
- [22] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 1, 2, 5, 6, 7
- [23] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *ICCV*, 2019. 1, 3
- [24] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 2
- [25] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017. 6
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 2015. 4
- [27] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015. 2, 4
- [28] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 1, 2
- [29] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, 2019. 2, 7
- [30] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 1
- [31] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *TPAMI*, 2012. 1
- [32] Yale Song, David Demirdjian, and Randall Davis. Continuous body and hand gesture recognition for natural human-computer interaction. *TIIS*, 2012. 1

- [33] Evan A Suma, Belinda Lange, Albert Skip Rizzo, David M Krum, and Mark Bolas. Faast: The flexible action and articulated skeleton toolkit. In *VRC*, 2011. 1
- [34] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 2
- [35] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2
- [36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 4
- [37] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, 2020. 2, 6, 7
- [38] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d pose estimation. In *NeurIPS*, 2021. 6
- [39] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *ICCV*, 2021. 3
- [40] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *ECCV*, 2020. 1, 3
- [41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2
- [42] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 3
- [43] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 3
- [44] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 6
- [45] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *NIPS*, 2018. 2, 6
- [46] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, 2021. 6
- [47] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, 2020. 1, 2, 6, 7
- [48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1, 3