Learning Structure-Guided Diffusion Model for 2D Human Pose Estimation

Zhongwei Qiu[1,3]   Qiansheng Yang[2]   Jian Wang[2]   Xiyu Wang[3]   Chang Xu[3]

Dongmei Fu[1]   Kun Yao[2]   Junyu Han[2]   Errui Ding[2]   Jingdong Wang[2]

[1] University of Science and Technology Beijing, [2] Baidu, [3] University of Sydney

Abstract

*One of the mainstream schemes for 2D human pose estimation (HPE) is learning keypoints heatmaps by a neural network. Existing methods typically improve the quality of heatmaps by customized architectures, such as high-resolution representation and vision Transformers. In this paper, we propose **DiffusionPose**, a new scheme that formulates 2D HPE as a keypoints heatmaps generation problem from noised heatmaps. During training, the keypoints are diffused to random distribution by adding noises and the diffusion model learns to recover ground-truth heatmaps from noised heatmaps with respect to conditions constructed by image feature. During inference, the diffusion model generates heatmaps from initialized heatmaps in a progressive denoising way. Moreover, we further explore improving the performance of DiffusionPose with conditions from human structural information. Extensive experiments show the prowess of our DiffusionPose, with improvements of 1.6, 1.2, and 1.2 mAP on widely-used COCO, CrowdPose, and AI Challenge datasets, respectively.*

1. Introduction

Human pose estimation is a fundamental computer vision task, which requires localizing the keypoints coordinates from the human body in an image and has many applications in action recognition [26], human reconstruction [55], and human neural radiance fields [49], etc. The modern approaches for 2D human pose estimation can be categorized as top-down [48, 40, 50], bottom-up [20, 7, 12], and one-stage methods [56, 37, 51]. Usually, the top-down methods show better performances than other paradigms due to the unified scale of persons and higher resolutions of people. And we focus on top-down manner in this paper.

Learning heatmaps is widely adopted in top-down [48, 40, 50] and bottom-up methods [20, 7] because it can reserve the spatial information and make the model easy to converge. To obtain better heatmap representation, existing methods typically design customized architectures to extract better features, such as high-resolution representa-
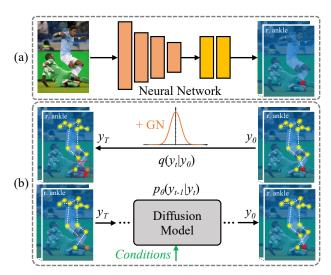

Figure 1. (a) Existing heatmap-based pose estimation framework [48]. (b) Diffusion-generation scheme, which includes a forward diffusion process q by adding Gaussian Noises (GN) and the reverse process p_θ of recovering heatmaps from noised heatmaps.

tion [40, 7] and vision Transformers [52, 24, 50]. For example, HRNet [40] maintains high-resolution features during each downsampling stage of the neural network. ViT-Pose [50] utilizes a full vision Transformer network to extract stronger feature representation. As shown in Figure 1 (a), these methods all follow a discriminative scheme of detecting each pixel of an image. Besides, some refinement-based methods [19, 54, 28] improve the accuracy of coordinates outputted from heatmaps by a two-stage refining model [28], post-process based on heatmap distribution [54], and learning keypoints offsets relative to heatmap peaks [19]. However, the performances of these post-process refinements are inevitably impacted by the quality of the original heatmaps. Our motivation is to explore a novel scheme for generating high-quality heatmaps.

Recently, diffusion models [16, 38] have shown superior performance in multi-modal content generation tasks, such as text-guided image/video generation [30, 17], image super-resolution [21, 36], 3D shapes and textures synthe-

sis [27], and so on. These successes are from the strong ability of the diffusion model that recovers the signals from noised inputs in a progressive way. Inspired by that, researchers successfully apply the diffusion denoising framework to object detection [4] and segmentation [14, 2, 1], which brings a promising prospect of tackling visual perception tasks via diffusion models. However, few explorations have been studied on applying diffusion models to 2D human pose estimation.

To generate high-quality heatmaps for 2D human pose estimation, we propose DiffusionPose, which formulates the 2D human pose estimation as a generative process of heatmaps from noised heatmaps via diffusion model. The novel diffusion-generation scheme for 2D human pose estimation is shown in Figure 1 (b), which includes a forward diffusion process $q(y_t|y_0)$ and a reverse process $p_\theta(y_{t-1}|y_t)$. During training, the keypoints are diffused to random distribution by adding Gaussian noises and the diffusion model learns to recover ground-truth heatmaps from noised heatmaps with respect to conditions constructed by image feature. During inference, the diffusion model generates high-quality heatmaps from initialized heatmaps in a progressive denoising way, with the guidance of conditions. Furthermore, we explore improving the performance of DiffusionPose by incorporating various human structural information as conditions and studying the influences of different heatmaps resolutions on DiffusionPose. These components lead DiffusionPose to achieve favorable results on three 2D pose estimation datasets.

Our main contributions can be summarized as follows:

- We propose DiffusionPose, which formulates the 2D human pose estimation as a denoising process from noised heatmaps via diffusion model. To the best of our knowledge, this is the first successful application of the diffusion model in 2D human pose estimation.

- We further propose the structure-guided diffusion decoder (SGDD) and high-resolution SGDD to reform the diffusion model for human pose estimation, which leads DiffusionPose to achieve better performance.

- Extensive experiments on COCO, CrowdPose, and AI Challenge datasets show that DiffusionPose achieves favorable results against the previous schemes.

2. Related Work

2.1. Human Pose Estimation

2D Human Pose Estimation. Modern 2D human pose estimation methods [22, 32, 40, 50] usually adopt neural networks to extract deep features, then estimate human poses from deep features. Therefore, related works mainly focus on the backbone network [48, 33, 44], pose representation [29, 22], and the framework of multi-person pose

estimation [37, 35]. Hourglass [29] stacks a multi-stage CNN network to extract features and regress 2D keypoints heatmaps to localize the coordinates of human joints. Based on the heatmap representation, SimpleBaseline [48], HR-Net [40], and ViTPose [50] propose different backbone networks to extract better features, respectively. Except for the heatmap-based representation, some regression-based methods [41, 22] also explore the different expressions of 2D human poses. To tackle the multiple instances in an image, three different frameworks are proposed. Top-down methods [48, 40, 50] first detect human bounding boxes, then apply the single-person pose models on the cropped images with human boxes. Bottom-up methods [3, 7] first estimate human poses, then assign them to different persons. One-stage methods [46, 37] directly regress multi-person poses from an image by keypoints offsets or pose tokens in a single stage.

Pose Refinement. Some methods [11, 23, 43, 54, 42] refine learned human poses by structure or distribution information. Graph-PCNN [43] adopts a two-stage graph pose refinement module to refine the human pose from sampled keypoints on heatmaps. PoseFix [28] proposes a model-agnostic general human pose refinement network to refine the estimated 2D pose from any pose model. DARK [54] proposes a distribution-aware coordinate representation to modulate 2D heatmaps. Although great improvements have been achieved through these approaches, their successes still rely on the quality of original heatmaps generated by the first stage. In this paper, we aim to propose a new scheme for generating heatmaps via diffusion models.

2.2. Diffusion Model

Diffusion Model for Generation Tasks. Recently, diffusion models [16, 38, 39] have achieved superior performance on the generation tasks, such as image/video synthesis [30, 17], image super-resolution [21, 36, 34], 3D shape and synthesis [27], etc. Diffusion models add noises to the signals and learn to recover signals from noised data samples by a progressive denoising process. Although diffusion achieves great success on many generation tasks, it has huge costs on computational resources. Thus, DDIM [38] and its variant [45] transform the denoising process to the non-Markov process from the Markov process by skipping-step sampling, which improves the inference speed and booms the application of diffusion models on other computer vision tasks.

Diffusion Model for Perception Tasks. With the success of diffusion models on generation tasks, researchers pay attention to applying diffusion models to other perception tasks, such as object detection [4], segmentation [14, 2, 1, 5], 3D pose lifting [13, 18], and so on. DiffusionDet [4] diffuses ground-truth object bounding boxes to noised boxes, then utilizes the diffusion model to re-

cover ground-truth bounding boxes from randomly sampled boxes. Based on DiffusionDet, DiffusionInst [14] adds a segmentation head on the regression head of Diffusion-Det, which achieves instance segmentation by the diffusion model. For 3D pose lifting, some works [13, 18] start from a randomly sampled 3D pose and recover the 3D pose with the guidance of 2D pose context. The success of diffusion models on perception tasks inspires us to study how to apply the diffusion model in 2D pose estimation and utilize its strong ability of recovering signals from noises to generate high-quality keypoints heatmaps.

3. Approach

3.1. Preliminary

Diffusion models achieve great success on content generation tasks [16, 17, 36], which includes the forward diffusion process and reverse denoising process. Given a data distribution y_0, the forward process q adds Gaussian noise to y_0 over T steps:

$$q(y_t|y_{t-1}) = \mathcal{N}(y_t|\sqrt{\alpha_t}y_{t-1}, (1-\alpha_t)\boldsymbol{I}), \qquad (1)$$

where $0 < \alpha_t < 1$ is the weight hyper-parameter, which controls the variance of the added noise. Particularly, given y_0, the sampled y_t can be generated as:

$$q(y_t|y_0) = \mathcal{N}(y_t|\sqrt{\gamma_t}y_0, (1-\gamma_t)\boldsymbol{I}), \qquad (2)$$

where $\gamma_t = \prod_{i=1}^{t} \alpha_i$. Practically, Equation 2 can be recapped as:

$$y_t = \sqrt{\gamma_t}y_0 + \sqrt{1-\gamma_t}\epsilon, \epsilon \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}). \qquad (3)$$

where ϵ represents the sampled noises.

Recently, many works [4, 14, 21] have applied the diffusion models to the perception tasks in computer vision. For object detection [4] and instance segmentation [14], a neural network $f_\theta(y_t, t, x)$ is trained to estimate y_0 from noised y_t during the training process. During the inference, the y_0 is estimated from noised y_t by diffusion model f_θ, conditioned on the corresponding image x.

In this work, we formulate the 2D human pose estimation as the task of 2D heatmap generation by the diffusion model. During training, the human keypoints y_0 are diffused to noised points y_t, which is further used to generate keypoints mask y_t^{mask}. Then, the feature condition x^c is extracted from image x based on y_t^{mask}. With the guidance of x^c, the diffusion model learns to recover ground-truth heatmaps y_0^{hm} from noised heatmaps y_t^{hm}. During inference, starting from noised heatmaps y_t^{hm}, the diffusion model estimates y_0^{hm} in a progressive way, which is eventually used to decode ground-truth keypoints y_0.

3.2. DiffusionPose

The framework of DiffusionPose is shown in Figure 2, which contains the forward diffusion process (FDP), the model forward process (MFP), and the reverse diffusion process (RDP). As shown in Figure 2, the blue flow lines, black flow lines, and red flow lines indicate the FDP, MFP, and RDP, respectively. During training, FDP diffuses ground-truth keypoints to noised keypoints by adding Gaussian noises, further generating noised heatmaps and feature masks. Then, MFP extracted features from the input image with feature masks as the conditions for the diffusion model. The decoder of DiffusionPose learns to recover ground-truth heatmaps from noised heatmaps with the guidance of conditions. During inference, RDP recovers the heatmaps from initialized heatmaps by diffusion decoder in a progressive denoising way and decodes the keypoints coordinates from estimated heatmaps.

3.2.1 Forward Diffusion Process

The Forward Diffusion Process (FDP) is shown in Figure 2 (Blue flow process) and Algorithm 1. Given ground-truth keypoints $y_0 \in \mathbb{R}^{J \times 2}$ and time step $t \in [1, T]$, FDP generates noised heatmaps $y_t^{hm} \in \mathbb{R}^{J \times H \times W}$ and feature masks y_t^{mask}. The diffusion process can be formulated as:

$$\begin{aligned} y_t &= q(y_t|y_0, \zeta) \\ &= \sqrt{\gamma_t}(\zeta \cdot y_0) + \sqrt{1-\gamma_t}\epsilon, \epsilon \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}), \end{aligned} \qquad (4)$$

where $q(\cdot)$ is the process of sampling in Equation 2 and 3. ζ is a scale parameter to control the ratio of signal and noise.

After sampling y_t, FDP generates noised heatmaps y_t^{hm} and feature masks y_t^{mask}, which can be formulated as:

$$\begin{aligned} y_t^{hm} &= \phi(y_t, \sigma), \\ y_t^{mask} &= \{M^{kps}, M^{ske}\} = \varphi(y_t, \delta_{kps}, \delta_{ske}), \end{aligned} \qquad (5)$$

where $\phi(\cdot)$ represents the operation of generating 2D Gaussian heatmaps with center points y_t and sigma parameter σ following [48]. $\varphi(\cdot)$ represents the operation of generating 2D keypoints masks M^{kps} and skeleton masks M^{ske} according to keypoints y_t and its corresponding skeleton. δ_{kps} and δ_{ske} control the widths of keypoints masks and skeleton masks.

3.2.2 Model Forward Process

After diffusing keypoints y_0 to noised keypoints y_t, diffusion model f_θ learns to recover keypoints heatmaps from noised heatmaps. f_θ includes encoder $E(\cdot)$, spatial-channel cross-attention (SC-CA) module $A(\cdot)$, and decoder $D(\cdot)$.

The Model Forward Process (MFP) is shown in Figure 2 (Black flow process) and Algorithm 1. Given an image
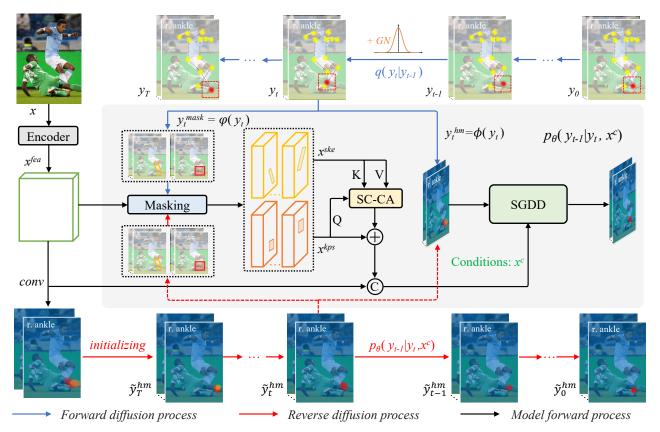
Figure 2. The framework of DiffusionPose, which includes the forward diffusion process (FDP), the model forward process (MFP), and the reverse diffusion process (RDP). During FDP, given step t, keypoints $y_0 \in \mathbb{R}^{J \times 2}$ diffuse to noised keypoints y_t by $q(\cdot)$ with Gaussian Noises (GN). y_t is used to generate mask y_t^{mask} by operation $\varphi(\cdot)$, which includes keypoints masks M^{kps} and skeleton masks M^{ske}. y_t is also used to generate the noised heatmaps y_t^{hm} by operation $\phi(\cdot)$. Based on y_t^{mask}, keypoints feature x^{kps} and skeleton feature x^{ske} are generated by masking x^{fea} from encoder. x^{kps} and x^{ske} are sent into SC-CA module to generate feature conditions x^c. The diffusion decoder (SGDD) learns to recover keypoints heatmaps from noised heatmaps. During RDP, SGDD estimates heatmaps from noised heatmaps in a progressive denoising way as $p_\theta(y_{t-1}|y_t, x^c)$. $\oplus$ and circled C are summing and concatenating operations, respectively.

x of size $H_x \times W_x$, noised heatmaps y_t^{hm}, feature masks M^{kps} and skeleton masks M^{ske}, DiffusionPose first extract deep feature by the encoder as $x^{fea} = E(x)$, where $E(\cdot)$ represents the encoder of diffusion model. Following previous work [40], HRNet pre-trained on ImageNet [8] dataset is used as the backbone network to extract features since it is the widely-used benchmark.

Then, keypoints features and skeleton features are extracted by the masking operation as follows:

$$x^{kps} = x^{fea} \otimes M^{kps}, x^{ske} = x^{fea} \otimes M^{ske}, \quad (6)$$

where $\otimes$ is element-wise multiplication. Next, the cross-attention is applied on x^{kps} and x^{ske} to capture structural information, and the output x^c is further used as the conditions of diffusion decoder, which can be formulated as:

$$x^c = C(x^{kps} \oplus A(x^{kps}, x^{ske}, x^{ske}), x^{fea}), \quad (7)$$

where $C(\cdot)$ and $\oplus$ represent concatenating features and summing, respectively. $A(\cdot)$ represents the SC-CA module

of DiffusionPose. SC-CA module aims to compute the relationship between keypoints features and skeleton features, which includes spatial attention and channel attention. The SC-CA module can be denoted as:

$$A(Q, K, V) = A_c(Q, K, V) + A_s(Q, K, V), \quad (8)$$

where A_c and A_s represent channel group attention and spatial window multi-head attention proposed in [9]. A_s aims to improve the accuracy of localization since the feature conditions x^c contain noises. A_c aims to enhance the semantic information of conditions. Q, K, and V are query, key, and value tokens, respectively.

After obtaining conditions x^c, the diffusion decoder $D(\cdot)$ estimates the keypoints heatmaps $\tilde{y}_{t-1}^{hm}$ as:

$$\tilde{y}_{t-1}^{hm} = D(y_t^{hm}, x^c). \quad (9)$$

For the diffusion decoder, we reform the diffusion model in content generation tasks and propose a Structure-Guided
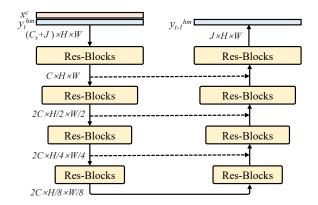
Figure 3. The architecture of Diffusion Decoder in DiffusionPose.

Stage	1	2	3	4
Res-Blocks	2	2	2	2
Channels	C	2C	2C	2C
Feature Size	$H \times W$	$H/2 \times W/2$	$H/4 \times W/4$	$H/8 \times W/8$

Table 1. The architecture parameters of SGDD. Channel dimension $C \in \{64, 128, 256\}$.

Diffusion Decoder (SGDD). To achieve better performance, we further propose high-resolution SGDD.

Structure-Guided Diffusion Decoder. The condition x^c with structural information is used to guide the generation of human heatmaps by SGDD. SGDD aims to recover the ground-truth heatmaps from noised heatmaps generated from the forward diffusion process, which contains a neural network with U-Net architecture as the previous works [16, 5]. As shown in Figure 3, given noised heatmap $y_t^{hm} \in \mathbb{R}^{J \times H \times W}$ and condition $x^c \in \mathbb{R}^{C_x \times H \times W}$, SGDD outputs the estimated heatmaps $y_0^{hm} \in \mathbb{R}^{J \times H \times W}$, where C_x, J, $H \times W$ represent the channel number of x^c, joints number, and the size of heatmaps, respectively.

In SGDD, input features and heatmaps are downsampled to a size of $2C \times \frac{H}{8} \times \frac{W}{8}$ in a progressive way, and up-sampled to the original size of $H \times W$ with the residual connection, where $C \in \{64, 128, 256\}$ is the channel number of features in the decoder. In each scale of SGDD, multiple Res-Blocks [15] are stacked to extract features and recover heatmaps. The architecture parameters of SGDD are shown in Table 1.

High-Resolution SGDD. To achieve better performance, SGDD can generate high-resolution heatmaps. As shown in Figure 3, the input size of the diffusion decoder is $H \times W$. In SGDD, $H = H_x/s$ and $W = W_x/s$, where $H_x \times W_x$ is the original size of input image x and $s \in \{1, 2, 4, 8\}$ is the scale of heatmaps. Thus, SGDD can run in choosing different scales. The diffusion decoding process with higher heatmap resolution could mitigate the quantization error and improve the quality of heatmaps.

Algorithm 1 Training DiffusionPose model f_θ

Input: image x, keypoints labels y_0, diffusion step T, f_θ contains encoder E and diffusion decoder D.

1: Extract feature $x^{fea} = E(x)$
2: **repeat**
3:     $t \sim \text{Uniform}(\{1, ..., T\})$
4:     $y_t = q(y_t|y_0, \zeta)$
5:     $y_t^{hm} = \phi(y_t, \sigma), y_0^{hm} = \phi(y_0, \sigma)$
6:     $y_t^{mask} = \{M^{kps}, M^{ske}\} = \varphi(y_t, \delta_{kps}, \delta_{ske})$
7:     $x^{kps} = x^{fea} \otimes M^{kps}, x^{ske} = x^{fea} \otimes M^{ske}$
8:     $x^c = C(x^{fea}, x^{kps} \oplus A(x^{kps}, x^{ske}, x^{ske}))$
9:     $L_1 = ||D(y_t^{hm}, x^c) - y_0^{hm}||_2$
10:     $L_2 = ||conv(x^{fea}) - y_0^{hm}||_2$
11:     Take gradient descent step on $\nabla_\theta(L_1 + L_2)$
12: **until** converged

Algorithm 2 Inference in T iterative steps

Input: image x, diffusion step T.

1: Extract feature $x^{fea} = E(x)$
2: $\tilde{y}_T^{hm} = conv(x^{fea})$
3: **for** $t = T, ..., 1$ **do**
4:     Deocde coordinates $\tilde{y}_t = argmax(\tilde{y}_t^{hm})$
5:     $y_t^{mask} = \{M^{kps}, M^{ske}\} = \varphi(\tilde{y}_t, \delta_{kps}, \delta_{ske})$
6:     $x^{kps} = x^{fea} \otimes M^{kps}, x^{ske} = x^{fea} \otimes M^{ske}$
7:     $x^c = C(x^{fea}, x^{kps} \oplus A(x^{kps}, x^{ske}, x^{ske}))$
8:     $\tilde{y}_{t-1}^{hm} = D(\tilde{y}_t^{hm}, x^c)$
9: **end for**
10: Deocde coordinates $\tilde{y}_0 = argmax(\tilde{y}_0^{hm})$
Return: $\tilde{y}_0$

3.2.3 Reverse Diffusion Process

During inference, starting from initialized heatmaps, DiffusionPose estimates the keypoints heatmaps in a progressive way. DiffusionPose adopts the strategy of skipping-step sampling as DDIM [38], which does not require inference steps equal to the training steps. The Reverse Diffusion Process (RDP) is shown in Figure 2 (red flow process) and Algorithm 2, which includes sampling initialization and reverse diffusion inference.

Sampling Initialization. A good sampling initialization could reduce the inference steps. In DiffusionPose, given the time step T, the noised heatmaps y_T are initialized as:

$$\tilde{y}_T^{hm} = conv(x^{fea}). \tag{10}$$

$\tilde{y}_T^{hm}$ can provide better initialization since it is trained with the supervision of ground-truth heatmaps.

Reverse Diffusion Inference. After initializing noised heatmaps, DiffusionPose estimates ground-truth heatmaps by SGDD. The process is shown in Algorithm 2. Feature masks y^{mask} are generated from $\tilde{y}_T^{hm}$, which is further used to generate conditions x^c as Equation 6 and 7. With $\tilde{y}_t^{hm}$

and x^c as inputs, SGDD estimates the $\tilde{y}_{t-1}^{hm}$ as Equation 9. Then, DiffusionPose iteratively estimates $\tilde{y}_0^{hm}$ by repeating the above steps. Finally, the keypoints $\tilde{y}_0$ are decoded from $\tilde{y}_0^{hm}$ by $argmax$ function.

3.3. Loss Function

DiffusionPose iteratively generates heatmaps from noised heatmaps. Following the previous works [48, 40, 50], L_2 loss is used as the loss function. To speed up the inference process, DiffusionPose learns coarse heatmaps from x^{fea} by a convolutional layer $conv$, which is used to initialize heatmaps for inference. Therefore, the total loss is

$$L = ||D(y_t^{hm}, x^c) - y_0^{hm}||_2 + ||conv(x^{fea}) - y_0^{hm}||_2. \quad (11)$$

4. Experiments

4.1. Implementation Details

The backbone network of DiffusionPose is HRNet [40] with pre-trained weights on ImageNet [8]. All ablation studies are based on HRNet-W32. During training, Adam optimizer with an initial learning rate of 5e-4 is used. DiffusionPose is trained by 210 epochs and the learning rate is divided by 10 at the 170th and 210th epochs. Data augmentations include random crop and random horizontal flip. DiffusionPose is trained on 8 Tesla A100 GPUs with a batch size of 32×8. During inference, the flip test is used following [40, 50]. Except specifically noted, the input size is 256×192.

4.2. Datasets and Evaluation Metric

We evaluate DiffusionPose on COCO [25], Crowd-Pose [23], and AI Challenge [47] datasets, respectively. All ablation studies are finished on the COCO dataset.

Datasets. COCO [25] dataset contains about 200k images and 250k person instances, in which each person is annotated with 17 keypoints. The training and validation subsets are used. For a fair comparison with [40, 50], the human bounding boxes from a detector with 56.4 mAP are used. CrowdPose [23] contains 20k images and 80k person instances, in which each person is annotated with 14 keypoints. Since more persons with serious occlusions are in an image, CrowdPose is more difficult. For a fair comparison with [53, 10], the ground-truth human bounding boxes when evaluating pose estimation models. AI Challenge [47] contains 350k images and 700k human instances, in which each person is annotated with 14 keypoints. For a fair comparison with [40, 53], the ground-truth human bounding boxes are used when testing pose estimation models.

Evaluation Metrics. Following the standard evaluation metrics of COCO datatset [25], the Object Keypoint Similarity (OKS) based metrics are used. In each dataset, AP, AP_{50}, AP_{75}, AR, AR_{50}, and AR_{75} are used. In the CrowdPose dataset, the APs on different crowd-index (easy, medium, hard) are also reported: AP_E, AP_M, and AP_H.

4.3. Comparisons with SOTA Methods

COCO dataset. We evaluate DiffusionPose on COCO dataset [25] and compare it with SOTA methods. As shown in Table 2, other methods are classified into four categories: Benchmarks, Refinement-based, and Transformer-based methods. Compared with HRNet benchmarks [40], HRNet-W32-based DiffusionPose achieves improvements of 1.5 and 1.3 AP with input sizes of 256×192 and 384×288, respectively. Compared with HRNet-W48, DiffusionPose also achieves improvements of 1.6 and 1.3 AP, respectively. The FLOPs of DiffusionPose are comparable with benchmarks. Compared with the refinement-based methods, which refine the keypoints heatmaps by learning keypoints offset [19], adjusting heatmap peaks [54], and extra models refining [28], DiffusionPose has the improvements of 0.4 to 1 AP. Compared with Transformer-based methods [24, 52, 53, 50], DiffusionPose shows improvements of 0.4 to 1.1 AP with the CNN backbone network.

CrowdPose dataset. To evaluate DiffusionPose on crowd scenarios, we test it on CrowdPose [23] dataset and compare it with SOTA methods. The images in the CrowdPose dataset contain multiple persons, which are more difficult to tackle since heavy occlusions. As shown in Table 3, DiffusionPose achieves new SOTA results of 76.7 and 77.5 AP with HRNet-W32 and HRNet-W48 as backbone networks, respectively. Compared with previous SOTA I^2R-Net [10] with TransPose [52] as backbone network, DiffusionPose has an improvement of 0.8 AP with a smaller HRNet-W48. Specifically, DiffusionPose achieves improvements of 1.2, 1.5, and 1.3 AP on AP_E, AP_M, and AP_H, respectively. These results show the stronger ability of DiffusionPose on handling occlusion cases in the crowd scenarios from the scheme of generating heatmaps.

AI Challenge dataset. We evaluate DiffusionPose on the challenging in-the-wild dataset: AI Challenge [47]. As shown in Tabel 4, DiffusionPose achieves state-of-the-art performances by 34.7 and 35.6 AP with HRNet-W32 and HRNet-W48 as the backbone networks, respectively. Compared with benchmark HRNet [40], DiffusionPose brings improvements of 2.4 and 2.1 AP on W32 and W48, respectively. Compared with HRFormer [53], DiffusionPose outperforms it by 1.2 AP. These results show the effectiveness and generalization ability of DiffusionPose.

4.4. Ablation Study

Main Components of SGDD. The ablation study of the main components of the SGDD is shown in Table 5. In Table 5, the default diffusion inference step is 1. The baseline of HRNet-W32 achieves 74.4 AP and 78.9 AR on COCO [25] dataset. DiffusionPose with image features as

#	Method	Backbone	Params (M)	FLOPs (G)	Resolution	Feature Scale	AP	AR
Benchmarks	SimpleBaseline [48]	ResNet-152	68	29	256x192	1/32	73.5	79.0
	HRNet [40]	HRNet-W32	29	8	256x192	1/4	74.4	78.9
	HRNet [40]	HRNet-W32	29	18	384x288	1/4	75.8	81.0
	HRNet [40]	HRNet-W48	64	16	256x192	1/4	75.1	80.4
	HRNet [40]	HRNet-W48	64	36	384x288	1/4	76.3	81.2
Refinement	DARK [54]	HRNet-W32	29	18	384x288	1/4	76.6	81.5
	DARK [54]	HRNet-W48	64	36	384x288	1/4	76.8	81.7
	UDP [19] + DARK [54]	HRNet-W32	29	16	384x288	1/4	76.8	81.5
	UDP [19] + DARK [54]	HRNet-W48	64	36	384x288	1/4	77.2	82.0
	PoseFix [28]	W48+Res-152	132	65	384x288	1/4	77.2	82.0
Transformer	TokenPose-L/D24[†] [24]	HRNet-W48	28	11	256x192	1/4	75.8	80.9
	TransPose-H/A6[†] [52]	HRNet-W48	18	22	256x192	1/4	75.8	80.8
	HRFormer-B[†] [53]	HRFormer-B	43	14	256x192	1/4	75.6	80.8
	HRFormer-B[†] [53]	HRFormer-B	43	29	384x288	1/4	77.2	82.0
	ViTPose-B(+UDP) [50]	ViT-B	86	22	256x192	1/16	75.8	81.1
Diffusion	**DiffusionPose**	HRNet-W32	38	14	256x192	1/4	75.9 (↑1.5)	81.1
	DiffusionPose	HRNet-W32	38	31	384x288	1/4	77.1 (↑1.3)	81.8
	DiffusionPose	HRNet-W48	74	22	256x192	1/4	76.7 (↑1.6)	81.7
	DiffusionPose	HRNet-W48	74	49	384x288	1/4	**77.6** (↑1.3)	**82.4**

Table 2. The comparison of DiffusionPose and SOTA methods on COCO [25] dataset. † indicates the Transformer-based methods. Other results are cited from [50]. All results are classified into four categories: Benchmarks, Refinement-based methods, Transformer-based methods, and DiffusionPose. ↑ shows the improvements compared with benchmarks.

Method	Backbone	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}	AP_E	AP_M	AP_H
OPEC-Net [31]	ResNet-101	70.6	86.8	75.6	-	-	-	-	-	-
HRNet [40]	HRNet-W32	71.3	91.1	77.5	-	-	-	80.5	71.4	62.5
TransPose-H[†] [52]	HRNet-W48	71.8	91.5	77.8	75.2	92.7	80.4	79.5	72.9	62.2
HRFormer-B[†] [53]	HRFormer-B	72.4	91.5	77.9	75.6	92.7	81.0	80.0	73.5	62.4
I²R-Net [10]	HRNet-W48	72.3	92.4	77.9	76.5	93.2	81.9	79.9	73.2	62.8
I²R-Net[†] [10]	TransPose-H	76.3	93.5	82.2	79.1	94.0	84.4	83.2	77.0	67.4
DiffusionPose	HRNet-W32	76.7	93.5	82.4	79.5	94.2	84.9	83.7	77.6	67.7
DiffusionPose	HRNet-W48	**77.5**	**93.6**	**83.5**	**80.3**	**94.5**	**85.6**	**84.4**	**78.5**	**68.7**

Table 3. The comparison of DiffusionPose and SOTA methods on CrowdPose [23] dataset. Other results are cited from [10].

Method	Backbone	AP	AP_{50}	AR	AR_{50}
SimpleBaseline [48]	ResNet-50	28.0	71.6	32.1	74.1
SimpleBaseline [48]	ResNet-101	29.4	73.6	33.7	76.3
SimpleBaseline [48]	ResNet-152	29.9	73.8	34.3	76.9
HRNet [40]	HRNet-W32	32.3	76.2	36.6	78.9
HRNet [40]	HRNet-W48	33.5	78.0	37.9	80.0
HRFormer[†] [53]	HRFomer-S	31.6	75.9	35.8	78.0
HRFormer[†] [53]	HRFomer-B	34.4	78.3	38.7	80.9
DiffusionPose	HRNet-W32	34.7	78.0	39.4	80.1
DiffusionPose	HRNet-W48	**35.6**	**79.1**	**40.1**	**81.1**

Table 4. Comparison with SOTA methods on AI Challenger [47].

Model	Diffusion	Conditions x^c	AP	AR
Baseline	×	−	74.4	78.9
DiffusionPose	✓	x^{fea}	75.0	80.0
DiffusionPose	✓	$C(x^{fea}, x^{kps})$	75.2	80.3
DiffusionPose	✓	$C(x^{fea}, A(x^{kps}, x^{ske}, x^{ske}))$	**75.4**	**80.5**

Table 5. The ablation study of main components of SGDD. $C(\cdot)$ denotes concatenating operation and A is the SC-CA module.

conditions achieves 75.0 AP. Adding keypoints features as conditions results in an improvement of 0.2 AP. The final version of DiffusionPose achieves 75.4 AP with improvements of 1 AP, with the SC-CA module to capture structural information from kepoints and skeleton features.

Diffusion Steps and Computation Costs. Diffusion-Pose iteratively generates high-quality keypoints heatmaps. Figure 4 shows the increasing performances and FLOPs of DiffusionPose with increasing inference steps. During inference, Diffusion initializes input heatmaps from coarse heatmaps by $conv(x^{fea})$, which is supervised ground-truth heatmaps and provides a good starting point for the diffusion model. As shown in Figure 4, just one reverse step achieves 75.4 AP. With more inference steps, the gains are about 0.1-0.2 AP and the FLOPs increase linearly. Thus, we
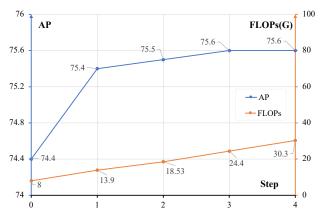
Figure 4. The performances and FLOPs of DiffusionPose with the increasing of sampling steps. [Best viewed in color]

Model	s	Resolution	AP	AR
DiffusionPose	8	32x24	72.2	78.1
	4	64x48	75.4	80.5
	2	128x96	75.9	81.1
	1	256x192	76.1	81.3

Table 6. The ablation study of SGDD on different resolutions.

Model	Backbone	Signal Scale ζ	AP	AR
DiffusionPose	HRNet-W32	0.01	75.3	80.5
		0.05	**75.4**	**80.5**
		0.10	75.2	80.4
		0.50	71.6	78.0
		1.00	70.7	77.5

Table 7. The ablation study of diffusion signal scale ζ.

suggest 1 step during inference for the trade-off of performance and cost. Compared with previous image/video generation models [16, 18] and detection/segmentation models [21, 14], DiffusionPose achieves better results with fewer inference steps from good initialized heatmaps y_T^{hm}.

High-resolution Diffusion Decoder. DiffusionPose supports generating heatmaps in different resolutions. Usually, high-resolution heatmaps result in better performance in 2D human pose estimation as [40, 53, 19]. We also conduct the ablation study of different resolutions of generated heatmaps. As shown in Table 6, DiffusionPose achieves better performances with a higher resolution. Considering the trade-off between computation costs and performance, we suggest adopting the setting of $s = 2$, which achieves 75.9 AP with the input image size of 256×192.

Signal Scale. In the diffusion process, hyper-parameter ζ controls the signal scale, which greatly affects the performance of DiffusionPose. As shown in Table 7, DiffusionPose achieves the best performance when $\zeta = 0.05$. Similar to segmentation task [5], a smaller scale factor of 0.05 is better than the standard scaling of 1.0 in previous work [6].

Figure 5. The joints heatmaps generated by traditional scheme [40] and our diffusion-generation scheme on the challenging cases.

Visualization Analysis. We compare the generated keypoints heatmaps by DiffusionPose and benchmark HRNet [40] in Figure 5. DiffusionPose with HRNet as backbone network achieves better pose results on these challenging cases from CrowdPose [23] dataset by generating better keypoints heatmaps.

4.5. Limitations and Discussion

Although DiffusionPose successfully introduces the diffusion models into 2D human pose estimation and achieves great results on three datasets, there are some limitations. 1) As shown in Figure 4, the FLOPs of DiffusionPose increase linearly with the inference steps, which limits the application of the diffusion model. We introduce an insight that sampling from a good initialization could alleviate this problem. How to generate content in one step is still worth exploring. 2) DiffusionPose extracts deep features by an image encoder as the conditions of the diffusion model. Directly using original images as the conditions is worth exploring, which could reduce the computation costs.

5. Conclusion

In this paper, we formulate 2D human pose estimation as a new scheme that generates heatmaps from noised heatmaps. We propose DiffusionPose, the first work to apply the diffusion model to 2D human pose estimation. To generate high-quality heatmaps, DiffusionPose diffuses keypoints to random distribution by adding noises and estimates 2D heatmaps from noised heatmaps in a progressive denoising way. Furthermore, we propose the structure-guided diffusion decoder (SGDD) and high-resolution SGDD for DiffusionPose, leading it to achieve significant improvements on three widely-used datasets.

References

[1] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *ICLR*, 2022. 2

[2] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *CVPR*, pages 4175–4186, 2022. 2

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[4] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 2, 3

[5] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*, 2022. 2, 5, 8

[6] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 8

[7] 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 *CVPR*, pages 5386–5395, 2020. 1, 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 4, 6

[9] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *ECCV*, pages 74–92. Springer, 2022. 4

[10] Yiwei Ding, Wenjin Deng, Yinglin Zheng, Pengfei Liu, Meihong Wang, Xuan Cheng, Jianmin Bao, Dong Chen, and Ming Zeng. I^{2}R-Net: Intra-and inter-human relation network for multi-person pose estimation. *IJCAI*, 2022. 6, 7

[11] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. In *CVPRW*, pages 205–214, 2018. 2

[12] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, pages 14676–14686, 2021. 1

[13] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. *arXiv preprint arXiv:2211.16940*, 2022. 2, 3

[14] Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, Jun Lan, Changhua Meng, and Weiqiang Wang. Diffusioninst: Diffusion model for instance segmentation. *arXiv preprint arXiv:2212.02773*, 2022. 2, 3, 8

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1, 2, 3, 5, 8

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 1, 2, 3

[18] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. *arXiv preprint arXiv:2211.16487*, 2022. 2, 3, 8

[19] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *CVPR*, pages 5700–5709, 2020. 1, 6, 7, 8

[20] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, pages 11977–11986, 2019. 1

[21] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1, 2, 3, 8

[22] 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*, pages 11025–11034, 2021. 2

[23] 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 *CVPR*, pages 10863–10872, 2019. 2, 6, 7, 8

[24] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *ICCV*, pages 11313–11322, 2021. 1, 6, 7

[25] 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*, pages 740–755. Springer, 2014. 6, 7

[26] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 1

[27] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 2

[28] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *CVPR*, pages 7773–7781, 2019. 1, 2, 6, 7

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 2

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804. PMLR, 2022. 1, 2

[31] 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 *ECCV*, pages 488–504. Springer, 2020. 7

9

[32] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Learning recurrent structure-guided attention network for multi-person pose estimation. In *ICME*, pages 418–423. IEEE, 2019. 2

[33] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. In *AAAI*, volume 34, pages 11924–11931, 2020. 2

[34] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *ECCV*, pages 257–273. Springer, 2022. 2

[35] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *CVPR*, pages 21254–21263, 2023. 2

[36] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 1, 2, 3

[37] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *CVPR*, pages 11069–11078, 2022. 1, 2

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1, 2, 5

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021. 2

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1, 2, 4, 6, 7, 8

[41] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, pages 2602–2611, 2017. 2

[42] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *CVPR*, pages 11060–11068, 2022. 2

[43] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-pcnn: Two stage human pose estimation with graph pose refinement. In *ECCV*, pages 492–508. Springer, 2020. 2

[44] Yunke Wang, Bo Du, and Chang Xu. Multi-tailed vision transformer for efficient inference. *arXiv preprint arXiv:2203.01587*, 2022. 2

[45] Yunke Wang, Xiyu Wang, Dung Dinh Anh, Bo Du, and Chang Xu. Learning to schedule in diffusion probablisitic models. In *KDD*, 2023. 2

[46] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *ECCV*, pages 527–544. Springer, 2020. 2

[47] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *ICME*, pages 1480–1485. IEEE, 2019. 6, 7

[48] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1, 2, 3, 6, 7

[49] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *NeurIPS*, 34:14955–14966, 2021. 1

[50] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 1, 2, 6, 7

[51] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *ICLR*, 2023. 1

[52] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *ICCV*, pages 11802–11812, 2021. 1, 6, 7

[53] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. In *NeurIPS*, 2021. 6, 7, 8

[54] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, pages 7093–7102, 2020. 1, 2, 6, 7

[55] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, pages 7739–7749, 2019. 1

[56] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1