2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.01356

Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation

Wenkang Shan[1], Zhenhua Liu[3], Xinfeng Zhang[2], Zhao Wang[1], Kai Han[3],
Shanshe Wang[1,4], Siwei Ma[1,4], Wen Gao[1,4]
[1]National Engineering Research Center of Visual Technology, Peking University
[2]University of Chinese Academy of Sciences
[3]Huawei Noah's Ark Lab [4]Peng Cheng Laboratory
{wkshan,zhaowang,sswang,swma,wgao}@pku.edu.cn, xfzhang@ucas.ac.cn,
{liu.zhenhua,kai.han}@huawei.com

Abstract

In this paper, a novel Diffusion-based 3D Pose estimation (D3DP) method with Joint-wise reProjection-based Multi-hypothesis Aggregation (JPMA) is proposed for probabilistic 3D human pose estimation. On the one hand, D3DP generates multiple possible 3D pose hypotheses for a single 2D observation. It gradually diffuses the ground truth 3D poses to a random distribution, and learns a denoiser conditioned on 2D keypoints to recover the uncontaminated 3D poses. The proposed D3DP is compatible with existing 3D pose estimators and supports users to balance efficiency and accuracy during inference through two customizable parameters. On the other hand, JPMA is proposed to assemble multiple hypotheses generated by D3DP into a single 3D pose for practical use. It reprojects 3D pose hypotheses to the 2D camera plane, selects the best hypothesis joint-by-joint based on the reprojection errors, and combines the selected joints into the final pose. The proposed JPMA conducts aggregation at the joint level and makes use of the 2D prior information, both of which have been overlooked by previous approaches. Extensive experiments on Human3.6M and MPI-INF-3DHP datasets show that our method outperforms the state-of-the-art deterministic and probabilistic approaches by 1.5% and 8.9%, respectively. Code is available at https://github.com/paTRICK-swk/D3DP.

1. Introduction

Monocular 3D human pose estimation aims to locate the 3D positions of human body joints from 2D images or videos. It plays a crucial role in various applications, such as human-computer interaction, metaverse, and self-driving. This task can be decomposed into two steps: first estimating the 2D locations of human joints using off-the-shelf 2D keypoint detectors (*e.g.*, CPN [11], OpenPose [6]); and then mapping these 2D locations to their corresponding 3D positions. In this work, we focus on the latter step, also known as the 2D-to-3D lifting process, following recent approaches[14, 42, 64, 47].

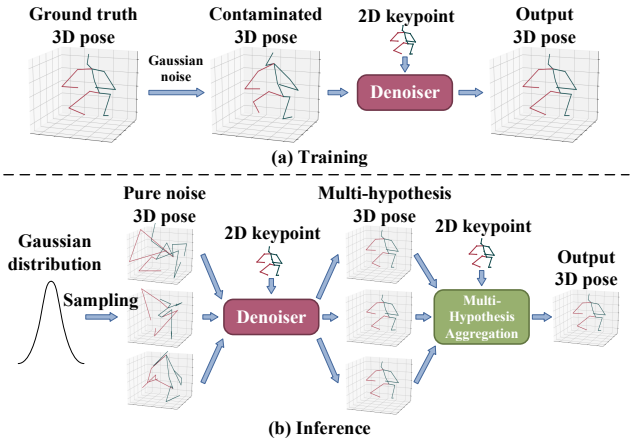

Figure 1: Thumbnail of our D3DP approach with multi-hypothesis aggregation. (a) Training: we gradually destroy the ground truth 3D pose and reverse this process by a denoiser to obtain a clean 3D prediction. (b) Inference: multiple Gaussian noises are passed through the denoiser, which is conditioned on 2D keypoints, to generate plausible 3D pose hypotheses. Then, the proposed multi-hypothesis aggregation method is utilized to yield the final output.

Existing works can be divided into two categories: *deterministic* and *probabilistic* approaches. Deterministic approaches [61, 36, 65, 47, 34] are designed to produce a single, definite 3D pose for each image, which is practical in real-world applications. Probabilistic approaches [59, 30, 24, 49, 40] represent the 2D-to-3D lifting as a probability distribution and produce a set of possible solutions for each image, which allows for uncertainty and ambiguity in the lifting process. This paper focuses on probabilistic meth-

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.01356

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

ods but combines the advantages of deterministic methods, *i.e.*, we aggregate multiple pose hypotheses into a single, higher-quality 3D pose for practical use.

Previous probabilistic approaches use generative models (*e.g.*, generative adversarial networks [17] or normalizing flows [43]) to predict multiple 3D pose hypotheses. Despite the great progress achieved, they still have the following two disadvantages: 1) They either (i) rely on special network designs and have poor compatibility [33, 59, 31, 52], or (ii) cannot specify the number of hypotheses based on actual needs [33, 30, 40]. 2) Although these methods generate multiple possible hypotheses, an individual 3D pose is still needed for practical use. They generally average over multiple hypotheses at the pose level to obtain the final output [49, 30], which does not take into account the differences between joints and the prior distribution of 2D keypoints. Consequently, the averaged result is much worse than the best of all hypotheses.

To address the first problem, we exploit Denoising Diffusion Probabilistic Models (DDPMs) [18] for 3D human pose estimation and introduce a novel approach called Diffusion-based 3D Pose estimation (D3DP). As shown in Fig. 1, our approach involves adding varying levels of noise to the ground truth 3D poses and learning a denoiser to predict the original data during training. During inference, we generate initial 3D poses by sampling pure noise from a Gaussian distribution. These poses are then sent to the denoiser, which is conditioned on 2D keypoints, to predict multiple 3D pose hypotheses. Compared to previous approaches, our method is superior in two aspects. (i) The denoiser has the compatibility to use existing 3D human pose estimators as the backbone, which enables an easy transformation from a deterministic approach into a probabilistic version. (ii) The number of hypotheses can be customized during inference. Besides, the proposed method progressively refines the generated results, which supports an additional adjustable parameter (number of iterations). More hypotheses and iterations bring better results at the expense of greater computational overhead. These two parameters help to strike a balance between performance and efficiency under limited computing resources.

To address the second problem, we focus on how to aggregate multiple 3D pose hypotheses into a single, more accurate 3D prediction. It is observed that the upper-bound performance of aggregation at the joint level is much higher than that at the pose level (Section 4.5), as the former conducts aggregation at a finer granularity. This observation drives us to propose a *joint-wise reprojection-based* multi-hypothesis aggregation (JPMA) method, which reprojects 3D pose hypotheses back to the 2D camera plane and conducts the joint-level selection of the best hypothesis based on the reprojection errors. The selected joints are assembled into a complete 3D pose as the final output. The proposed method uses *joints* as granularity, allowing a more diverse combination of hypotheses. In addition, the marginal distribution of 2D keypoints is taken into account, thus aiding the modeling of the posterior distribution of the final 3D pose. Compared to previous methods that use the averaged pose, JPMA integrates joint-level geometric priors into the multi-hypothesis aggregation, and hence improves the prediction performance.

Our contributions can be summarized as follows:

- We propose a diffusion-based 3D human pose estimation (D3DP) method, which is compatible and customizable.
- When multiple hypotheses are combined into a single prediction, we observe that the upper bound of performance for aggregation at the joint level is significantly higher than that at the pose level. This observation drives us to conduct joint-by-joint aggregation. To the best of our knowledge, we are the first to tackle this problem from the *joint* perspective.
- We propose a joint-wise reprojection-based multi-hypothesis aggregation (JPMA) method, which leverages the 2D prior at the joint level to improve the accuracy of the final 3D prediction. The proposed method proves to be more effective than existing pose-level aggregation methods.
- Our method outperforms both the state-of-the-art deterministic and probabilistic approaches on 3D human pose estimation benchmarks.

2. Related Work

2.1. Diffusion Model

Diffusion models [51] are a family of generative models that gradually deconstruct observed data by adding noise, and then recover the original data by reversing this process. Denoising Diffusion Probabilistic Models (DDPMs) [18] establish a connection between diffusion models and denoising score matching, which sparks recent interest in this approach. DDPMs have shown remarkable progress in numerous areas, including image generation [19, 38, 3, 44, 45, 46], cross-modal generation [1, 15, 22, 26, 29], graph generation [55, 39, 25, 60], semantic segmentation [2, 5], object detection [8], etc. For example, SR3 [46] adapts DDPMs to super-resolution. DiffusionDet [8] applies diffusion models to object detection to recover bounding boxes from noise. Pix2Seq-D [10] leverages diffusion models based on analog bits for panoptic segmentation.

3D human pose estimation suffers from intrinsic depth ambiguity, making probabilistic generation methods suitable for this task. Since diffusion models are well known for their ability to generate high-fidelity samples, we apply DDPMs to 3D human pose estimation.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

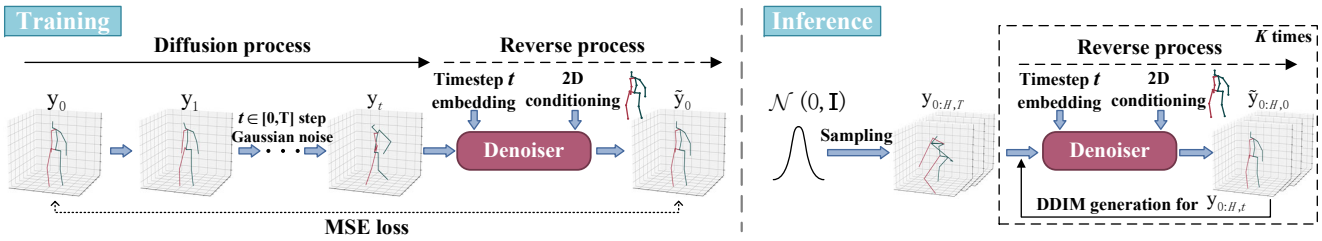

Figure 2: Overview of the proposed D3DP method. Left: Training. t-step Gaussian noise is added to the ground truth 3D pose $\boldsymbol{y}_0$, resulting in the noisy pose $\boldsymbol{y}_t$. $\boldsymbol{y}_t$ is then fed to a denoiser conditioned on 2D keypoints $\boldsymbol{x}$ and timestep t to yield the final prediction $\widetilde{\boldsymbol{y}}_0$. Right: Inference. H samples are drawn from a Gaussian distribution to initialize 3D poses $\boldsymbol{y}_{0:H,T}$, which are utilized to yield the noiseless 3D pose hypotheses $\widetilde{\boldsymbol{y}}_{0:H,0}$. Besides, we can iterate the above reverse process K times to refine the final results by sending DDIM-generated 3D poses $\boldsymbol{y}_{0:H,t}$ with different levels of noise to the denoiser.

2.2. Probabilistic 3D Human Pose Estimation

Early probabilistic methods [50, 28] rely on heuristic approaches to sample multiple plausible 3D poses that could potentially produce similar 2D projections. Recently, approaches based on mixture density networks (MDN) [4], normalizing flows (NF) [43], conditional variational autoencoder (CVAE) [52] and generative adversarial networks (GAN) [17] have been proposed. Li and Lee [30] introduce MDN to generate pose hypotheses, and Oikarinen *et al.* [40] further combine MDN with graph neural networks that incorporate graph-structured information. Sharma *et al.* [49] use CVAE to obtain diverse 3D pose samples. Wehrbein *et al.* [59] and Li *et al.* [31] exploit NF and GAN for this task, respectively. Besides, MHFormer [33] uses a Transformer [54] to generate three features to represent three individual hypotheses.

To overcome the drawbacks of previous methods, such as bad compatibility (MHFormer, NF, GAN, MDN, CVAE), and a non-adjustable number of hypotheses (MHFormer, MDN), we present D3DP, a straightforward and compatible architecture that can be easily integrated with existing 3D pose estimators with minimal modifications. Our method supports a custom number of iterations K and hypotheses H during inference to balance performance and efficiency. Note that concurrent methods [20, 16, 13] also use diffusion models for this task, but they either don't use any aggregation method or use a pose-level method (average). In contrast, we propose a joint-level aggregation method, called JPMA, to achieve a more accurate prediction.

2.3. Multi-Hypothesis Aggregation

Probabilistic 3D pose estimation methods typically generate multiple hypotheses, but how to synthesize them into a single 3D pose is still an open research question. Many methods [59, 30, 40, 31] simply average all pose hypotheses, which often produces unsatisfactory results. Other approaches explore more sophisticated ways of aggregating multiple hypotheses. Sharma *et al.* [49] use depth ordinal relations to score and weight-average the candidate 3D poses. Oikarinen *et al.* [40] choose the 3D pose with the highest mixture coefficient outputted by MDN. However, these works ignore the differences between joints. Besides, the information contained in 2D keypoints is not exploited.

Thus, our JPMA method is proposed to use the errors between the reprojected hypotheses and the input 2D keypoints as guidance for joint-level hypothesis selection. Although other approaches [42, 57, 12, 7, 27] have already used the 3D-to-2D reprojection method in deterministic 3D human pose estimation, our method marks the pioneering effort to apply it to probabilistic estimation. Specifically, reprojection is incorporated into the multi-hypothesis aggregation (MHA), which provides a new perspective on this problem.

3. Method

Given the input 2D keypoints $\boldsymbol{x} = \{p_n^{2\mathrm{d}}\}_{n=0}^{N}, p_n^{2\mathrm{d}} \in \mathbb{R}^{J\times 2}$, our goal is to predict the corresponding 3D positions of all joints $\widetilde{\boldsymbol{y}} = \{p_n^{3\mathrm{d}}\}_{n=0}^{N}, p_n^{3\mathrm{d}} \in \mathbb{R}^{J\times 3}$. N, J are the number of frames and human joints in each frame, respectively. Since diffusion models can generate high-quality samples, we apply it to probabilistic 3D human pose estimation and propose D3DP method, which generates multiple 3D pose hypotheses for a single 2D observation. Then, these hypotheses are used by the proposed JPMA method to yield a single, accurate 3D pose for practical use.

3.1. Diffusion-Based 3D Pose Estimation (D3DP)

D3DP uses two Markov chains: 1) a *diffusion process* that gradually perturbs data to noise, and 2) a *reverse process* that reconstructs the uncontaminated data by a denoiser. For more information about diffusion models, please refer to Appendix A. The training and inference phases of D3DP are described in detail below.

Training. As shown in Fig. 2 (left), we first sample a timestep $t \sim U(0, T)$, where T is the maximum number of timesteps. Then the ground truth 3D pose $\boldsymbol{y}_0$ is diffused

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

to the corrupted pose $\boldsymbol{y}_t$ by adding t-step independent Gaussian noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$. Following DDPMs [18], this process can be formulated as

$$q(\boldsymbol{y}_t \mid \boldsymbol{y}_0) := \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \epsilon\sqrt{1-\bar{\alpha}_t} \tag{1}$$

where $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$. β_t is the cosine noise variance schedule. When T is large enough, the distribution of $q(\boldsymbol{y}_T)$ is nearly an isotropic Gaussian distribution.

Subsequently, $\boldsymbol{y}_t$ is sent to a denoiser $\mathfrak{D}$ conditioned on 2D keypoints $\boldsymbol{x}$ and timestep t to reconstruct the 3D pose $\widetilde{\boldsymbol{y}}_0$ without noise:

$$\widetilde{\boldsymbol{y}}_0 = \mathfrak{D}\left(\boldsymbol{y}_t, \boldsymbol{x}, t\right) \tag{2}$$

The entire framework is supervised by a simple MSE loss:

$$\mathcal{L} = ||\boldsymbol{y}_0 - \widetilde{\boldsymbol{y}}_0||_2 \tag{3}$$

Inference. Since the degraded data is well approximated by a Gaussian distribution after the diffusion process, we can obtain H initial 3D poses $\boldsymbol{y}_{0:H,T}$ by sampling noise from a unit Gaussian. As shown in Fig. 2 (right), feasible 3D pose hypotheses $\widetilde{\boldsymbol{y}}_{0:H,0}$ are predicted by passing $\boldsymbol{y}_{0:H,T}$ to the denoiser $\mathfrak{D}$ (Eq. 2).

Thereafter, $\widetilde{\boldsymbol{y}}_{0:H,0}$ are used to generate the noisy 3D poses $\widetilde{\boldsymbol{y}}_{0:H,t'}$ as inputs to the denoiser for the next timestep via DDIM [53], which can be formulated as

$$\boldsymbol{y}_{0:H,t'} = \sqrt{\bar{\alpha}_{t'}} \cdot \widetilde{\boldsymbol{y}}_{0:H,0} + \sqrt{1-\bar{\alpha}_{t'}-\sigma_t^2} \cdot \epsilon_t + \sigma_t \epsilon \tag{4}$$

where t, t' are the current and next timesteps, respectively. The initial $t = T$. $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ is standard Gaussian noise independent of $\boldsymbol{y}_{0:H,t}$ and

$$\begin{aligned} \epsilon_t &= \left(\boldsymbol{y}_{0:H,t} - \sqrt{\bar{\alpha}_t} \cdot \widetilde{\boldsymbol{y}}_{0:H,0}\right) / \sqrt{1-\bar{\alpha}_t} \\ \sigma_t &= \sqrt{\left(1-\bar{\alpha}_{t'}\right) / \left(1-\bar{\alpha}_t\right)} \cdot \sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t'}} \end{aligned} \tag{5}$$

where ϵ_t is the noise at timestep t (derived from Eq. 1). σ_t controls how stochastic the diffusion process is.

Then, we can regenerate $\widetilde{\boldsymbol{y}}_{0:H,0}$ using $\boldsymbol{y}_{0:H,t'}$ as inputs to the denoiser. This procedure will be iterated K times. Since we start from T at the beginning, the timestep of each iteration can be written as $t = T \cdot (1 - k/K), k \in [0, K)$. The detailed algorithm is provided in Appendix C.

In this way, the proposed method supports a customizable number of hypotheses H by sampling multiple times from a Gaussian distribution. Moreover, the denoiser is trained only once but can be used multiple times to progressively refine the final predictions during inference. This refinement process supports an additional adjustable parameter (number of iterations K), which controls the diversity and quality of the generated hypotheses. Thus, we can specify an arbitrary number of H, K during inference (both are set to 1 during training), which allows users to balance performance and efficiency. This property solves the problem of a non-adjustable number of hypotheses faced by previous methods [33, 30, 40].

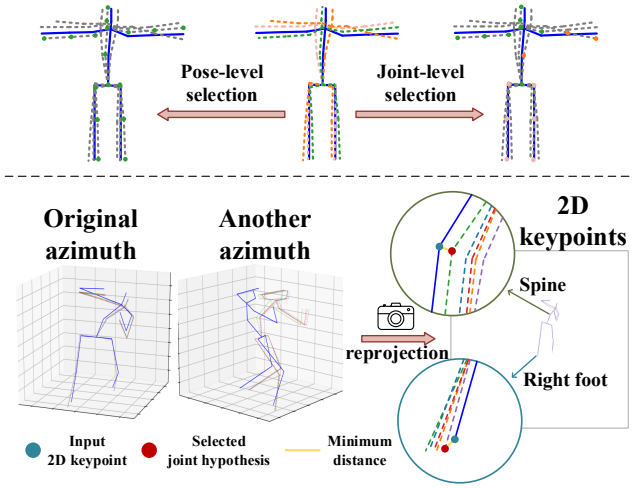

Figure 3: Top: Comparison between the pose-level (existing methods) and joint-level (proposed method) selections. The selected joint is drawn in the color of the hypothesis it belongs to. Bottom: Our JPMA method. 3D pose hypotheses are reprojected to the 2D camera plane and compared with the input 2D keypoints. For each joint, we select the best hypothesis that has the minimum distance to the 2D input. Dashed line: predicted 3D pose hypotheses. Each color represents an individual hypothesis. Solid blue line: ground truth 3D poses.

3.2. Joint-Wise Reprojection-Based Multi-Hypothesis Aggregation (JPMA)

Previous probabilistic 3D pose estimation methods [24, 43, 30] mainly focus on the generation of 3D pose hypotheses, with little attention to how to aggregate multiple hypotheses to produce a single, high-fidelity prediction. These approaches typically report the error of the best 3D pose (*i.e.*, closest to the ground truth) to highlight the superiority of their methods. However, this metric only demonstrates the upper bound of the performance and is not feasible in real-world applications because the ground truth is not available. Therefore, the averaged 3D pose is used as the final prediction, while its performance is far worse than the best result. A few works [40, 49] explore multi-hypothesis aggregation methods that are more effective than average for testing on in-the-wild videos. However, they use *poses* as granularity and ignore the distinctions between joints. Also, they do not establish the correspondence between 3D predictions and the 2D observation. Hence, further improvements are needed to achieve more accurate and reliable results.

To alleviate these problems, we first validate the upper-bound performance of aggregation methods at two levels by choosing the best hypothesis that is closest to the ground truth. In the case of pose level, the best pose is selected as the final prediction. In the case of joint level, the best

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

joint is selected and all the selected joints are combined into the final prediction. Experimental results show that the latter has a higher upper-bound performance than the former (Section 4.5), as the joint-level aggregation allows more flexible combinations of hypotheses. An intuitive comparison between the upper bounds of joint-level and pose-level selections (aggregations) is shown in Fig. 3 (top). When the pose-level selection is used, the green hypothesis (best pose) will be used as the results for all joints. However, the performance of the same hypothesis may vary across joints. Thus, the joint-level selection decomposes and recombines the hypotheses at the joint level instead of viewing each 3D pose as a whole. Only a fraction of joints (those with better performance) in each hypothesis will be used by the joint-level selection, and then these selected joints are assembled into an independent 3D pose. In this way, the advantage of each hypothesis is maximized. Moreover, the input 2D keypoints can be used as geometric priors to guide the model to choose the most likely 3D pose $\widetilde{\boldsymbol{y}}_0$ with the maximum posterior probability $q(\widetilde{\boldsymbol{y}}_0|\boldsymbol{x}, \widetilde{\boldsymbol{y}}_{0:H,0})$ as the final prediction during inference. Although 2D keypoints contain no depth information about 3D poses, they indicate the possible locations of human joints in 3D space (*i.e.*, the joint should be on the ray from the camera's optical center to the 2D keypoint). Thus, 2D keypoints still play an important role in multi-hypothesis aggregation.

Based on the above two ideas, we propose a joint-wise reprojection-based multi-hypothesis aggregation (JPMA) method, which is illustrated in Fig. 3 (bottom). We use known or estimated intrinsic camera parameters (4 for pinhole camera and 9 for distorted pinhole camera) to reproject 3D hypotheses $\widetilde{\boldsymbol{y}}_{0:H,0}$ to the 2D camera plane. Then, we calculate the distance between each hypothesis and the input 2D keypoint $\boldsymbol{x}$ for each joint, and select the joint hypothesis with the minimum distance (reprojection error). This process allows us to choose distinct hypotheses for different joints (green hypothesis for the spine and purple hypothesis for the right foot) and combine all the selected joints into the final prediction $\widetilde{\boldsymbol{y}}_0$, which can be formulated as:

$$\widetilde{\boldsymbol{y}}_0^{(i)} = \widetilde{\boldsymbol{y}}_{h',0}^{(i)}, \quad h' = \underset{h\in[0,H]}{\arg\min}||\mathcal{P}\left(\widetilde{\boldsymbol{y}}_{h,0}^{(i)}\right) - \boldsymbol{x}^{(i)}||_2 \quad (6)$$

where i is the index of joints. $\mathcal{P}(\cdot)$ is the reprojection function, which is detailed in Appendix B.

3.3. Architecture

The previous approaches [59, 33, 31, 52] rely on special network designs, for example, MHFormer [33] needs to use the cross-hypothesis interaction module and NF [59] requires the construction of bijective transformations. Therefore, they cannot keep up with the latest 3D human pose estimation methods. While the proposed D3DP method is compatible with up-to-date deterministic 3D pose estimators by using them as the backbone of the denoiser. In this paper, we use MixSTE [64], a mixed spatial temporal Transformer-based method, as the backbone. We can easily transform a deterministic pose estimation approach into a probabilistic version by making minor modifications, which are introduced in the next two paragraphs.

2D conditioning. Since our goal is to estimate 3D poses from 2D keypoints, using only noisy 3D poses $\boldsymbol{y}_{0:H,t}$ as inputs to the denoiser is not sufficient. Therefore, 2D keypoints $\boldsymbol{x}$ are used as additional information to guide the denoising process. We implement several schemes for fusing 2D keypoints and noisy 3D poses, such as cross attention, concatenation, etc. According to the experimental results (Appendix D.1), we choose to directly concatenate $\boldsymbol{x}$ and $\boldsymbol{y}_{0:H,t}$, and send them to the denoiser as inputs.

Timestep embedding. As the denoiser needs to handle 3D poses with different levels of noise, it is necessary to provide information about the current timestep t, which represents the number of times the Gaussian noise is added. Following DDPMs, we use a sinusoidal function to transform t into a timestep embedding and add it after the input embedding in the same way as the Transformer positional embedding [54].

These two modifications bring a negligible amount of computation. In addition, JPMA mainly consists of a reprojection process (without network parameters), so its amount of computation can also be ignored. Therefore, the computational overhead of the proposed D3DP with JPMA (when H=K=1) is essentially the same as that of the backbone network.

4. Experiments

4.1. Datasets and Evaluation Metrics

Human3.6M (H36M) [23] is the largest and most commonly used indoor dataset. It contains 15 activities (*e.g.*, walking) performed by 11 actors. Videos are captured by 4 synchronized and calibrated cameras at 50Hz. Following [33, 42, 40], our model is trained on 5 subjects (S1, S5, S6, S7, S8) and evaluated on 2 subjects (S9, S11). For evaluation metrics, we report the mean per joint position error (MPJPE), which computes the mean Euclidean distance between estimated and ground truth 3D joint positions in millimeters. Besides, Procrustes MPJPE (P-MPJPE) is also reported, which computes MPJPE after the estimated poses align to the ground truth using a rigid transformation.

MPI-INF-3DHP (3DHP) [37] is a recently popular dataset consisting of indoor and outdoor scenes. The training set contains 8 activities performed by 8 actors, and the test set covers 7 activities. Following [47], we use the valid frames provided by the official for testing. For evaluation metrics, we report MPJPE, percentage of correct keypoint (PCK)

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

Deterministic Methods																	
MPJPE		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
TCN [42] (N=243)*	CVPR'19	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
SRNet [61] (N=243)*	ECCV'20	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
PoseFormer [65] (N=81)*	ICCV'21	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
RIE [48] (N=243)*	MM'21	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.3
Anatomy [9] (N=243)*	TCSVT'21	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
U-CDGCN [21] (N=96)*	MM'21	38.0	43.3	39.1	39.4	45.8	53.6	41.4	41.4	55.5	61.9	44.6	41.9	44.5	31.6	29.4	43.4
Ray3D [62] (N=9)*	CVPR'22	44.7	48.7	48.7	48.4	51.0	59.9	46.8	46.9	58.7	61.7	50.2	46.4	51.5	38.6	41.8	49.7
STE [32] (N=351)*	TMM'22	39.9	43.4	40.0	40.9	46.4	50.6	42.1	39.8	55.8	61.6	44.9	43.3	44.9	29.9	30.3	43.6
P-STMO [47] (N=243)*	ECCV'22	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
MixSTE [64] (N=243)*‡	CVPR'22	37.9	40.1	37.5	39.4	43.3	50.0	39.8	39.9	52.5	56.6	42.4	40.1	40.5	27.6	27.7	41.0
MixSTE [64] (N=243)*	CVPR'22	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
DUE [63] (N=300)*	MM'22	37.9	41.9	36.8	39.5	40.8	49.2	40.1	40.7	47.9	53.3	40.2	41.1	40.3	30.8	28.6	40.6
D3DP (N=243, H=1, K=1)*		37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0

Probabilistic Methods																	
MPJPE		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
CVAE [49] (N=1, H=200, P-Agg)	ICCV'19	48.6	54.5	54.2	55.7	62.6	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
GAN [31] (N=1, H=10, P-Agg)	BMVC'20	67.9	75.5	71.8	81.8	81.4	93.7	75.2	81.3	88.8	114.1	75.9	79.1	83.3	74.3	79.0	81.1
GraphMDN [40] (N=1, H=5, P-Agg)	IJCNN'21	51.9	56.1	55.3	58.0	63.5	75.1	53.3	56.5	69.4	92.7	60.1	58.0	65.5	49.8	53.6	61.3
NF [59] (N=1, H=1, P-Agg)	ICCV'21	52.4	60.2	57.8	57.4	65.7	74.1	56.2	59.1	69.3	78.0	61.2	63.7	67.0	50.0	54.9	61.8
MHFormer [33] (N=351, H=3, P-Agg)*	CVPR'22	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
D3DP (N=243, H=1, K=1, P-Agg)*		37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0
D3DP (N=243, H=20, K=10, P-Agg)*		37.6	39.7	35.8	38.0	41.7	48.7	39.4	38.2	50.3	53.3	41.4	39.4	39.7	27.4	27.3	39.9
D3DP (N=243, H=20, K=10, J-Agg)*		37.3	39.5	35.6	37.8	41.3	48.2	39.1	37.6	49.9	52.8	41.2	39.2	39.4	27.2	27.1	39.5
MDN [30] (N=1, H=5, P-Best♯)	CVPR'19	43.8	48.6	49.1	49.8	57.6	64.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
CVAE [49] (N=1, H=200, P-Best♯)	ICCV'19	37.8	43.2	43.0	44.3	51.1	57.0	39.7	43.0	56.3	64.0	48.1	45.4	50.4	37.9	39.9	46.8
GAN [31] (N=1, H=10, P-Best♯)	BMVC'20	62.0	69.7	64.3	73.6	75.1	84.8	68.7	75.0	81.2	104.3	70.2	72.0	75.0	67.0	69.0	73.9
GraphMDN [40] (N=1, H=200, P-Best♯)	IJCNN'21	40.0	43.2	41.0	43.4	50.0	53.6	40.1	41.4	52.6	67.3	48.1	44.2	49.0	39.5	40.2	46.2
NF [59] (N=1, H=200, P-Best♯)	ICCV'21	38.5	42.5	39.9	41.7	46.5	51.6	39.9	40.8	49.5	56.8	45.3	46.4	46.8	37.8	40.4	44.3
D3DP (N=243, H=1, K=1, P-Best♯)*		37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0
D3DP (N=243, H=20, K=10, P-Best♯)*		37.3	39.4	35.4	37.8	41.3	48.1	39.0	37.9	49.8	52.8	41.1	39.0	39.4	27.3	27.2	39.5
D3DP (N=243, H=20, K=10, J-Best♯)*		33.0	34.8	31.7	33.1	37.5	43.7	34.8	33.6	45.7	47.8	37.0	35.0	35.0	24.3	24.1	35.4

Table 1: Results on Human3.6M in millimeters under MPJPE. N, H, K: the number of input frames, hypotheses, and iterations of the proposed D3DP. (‡) - Our implementation. (♯) - Not feasible in real-world applications. (*) - Use CPN [11] as the 2D keypoint detector to generate the inputs. Red: Best. Blue: Second best. Gray : our method.

within 150mm range, and area under curve (AUC).

3DPW [56] is the first dataset in the wild that includes video footage taken from a moving phone camera. We train our method on Human3.6M and evaluate on this dataset to show the qualitative results.

4.2. Implementation Details

We use MixSTE [64] as the backbone of the denoiser, which makes use of factorized attention by combining multiple spatial and temporal attention layers. It contains 16 alternating spatial and temporal Transformer layers [54] with the channel size set to 512. We modify MixSTE to fit the diffusion framework as follows: concatenating 2D keypoints and noisy 3D poses as inputs; adding timestep embedding after the input embedding in the first layer. The detailed architecture of the denoiser is shown in Fig. 4.

The proposed method is implemented in PyTorch [41] using AdamW [35] optimizer with the momentum parameters as $\beta_1, \beta_2 = 0.9, 0.999$ and the weight decay as 0.1. We train our model for 400 epochs and the initial learning rate is $6e^{-5}$ with a shrink factor of 0.993 after each epoch. We set the batch size to 4 and each sample contains a pose sequence of 243 frames (N=243). During training, the number of hypotheses and iterations H, K are set to 1,1, respectively. During inference, they are set to 20, 10. The maximum number of timesteps T is set to 1000. All experiments are carried out on two GeForce RTX 3080 Ti GPUs.

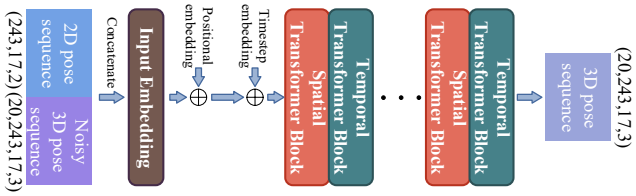

Figure 4: Detailed architecture of the denoiser. The tensor sizes are shown in parentheses. The 2D pose sequence (243,17,2) is expanded, repeated, and then concatenated with the noisy 3D pose sequence (20,243,17,3) to get the input sequence (20,243,17,5). 20, 243, 17 are the number of hypotheses H, frames N, and human joints J in each frame, respectively.

4.3. Quantitative Results

Human3.6M. As shown in Table 1 (top), we compare the proposed D3DP with the state-of-the-art deterministic 3D human pose estimation methods on Human3.6M. For fairness, we set H=1 to produce a single prediction. Our D3DP (N=243) achieves 40.0mm under MPJPE, outperforming DUE [63] (N=300) by 0.6mm, even though we use fewer frames. Moreover, our method exceeds MixSTE [64] (our backbone) by 1 mm under MPJPE. Although our method aims to generate multiple 3D pose hypotheses, it still improves performance in the single-hypothesis case. This result demonstrates that D3DP (*i.e.*, conditional denoising from corrupted 3D poses) helps the network generalize bet-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

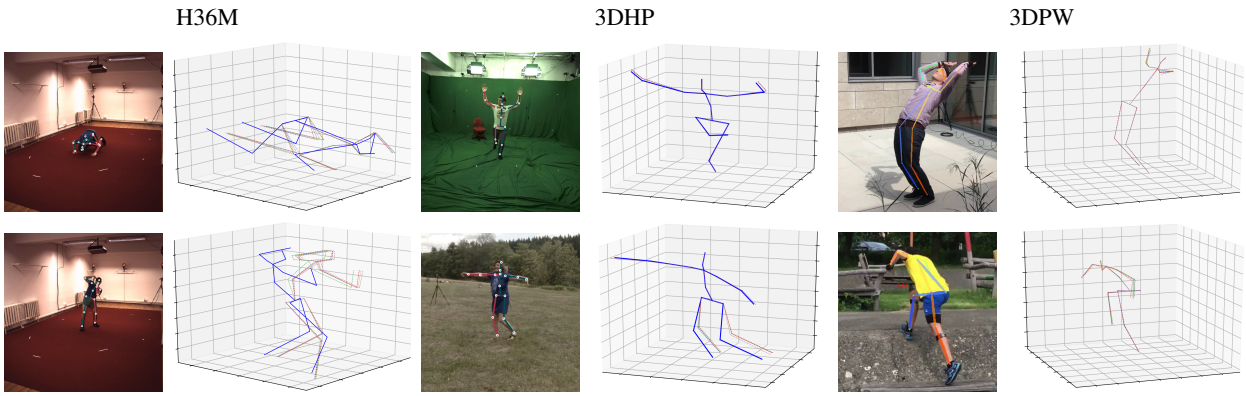

Figure 5: Qualitative results on three datasets. Dashed line: predicted 3D pose hypotheses. Each color represents an individual hypothesis. Solid blue line: ground truth 3D poses.

ter to the test set, thus improving the prediction accuracy.

Additionally, D3DP is compared with other probabilistic methods, as shown in Table 1 (bottom). We report the results according to the following four experimental settings.

- **P-Agg**: Use pose-level aggregation methods, which treat each 3D pose as the smallest unit to compose the final prediction, such as depth ordinal guidance [49], mean [40] and mean-shift [31].
- **J-Agg** (proposed): Use joint-level aggregation methods (our JPMA) to yield the final prediction. Each joint is treated as the smallest unit.
- **P-Best**♯: Select the 3D pose hypothesis that is closest to the ground truth. It is the upper bound of P-Agg performance.
- **J-Best**♯ (proposed): Select the joint hypothesis that is closest to the ground truth, and then combine the selected joints into the final 3D pose. It is the upper bound of J-Agg performance.

(♯) refers to the settings that are not feasible for in-the-wild videos. When testing under traditional P-Best and P-Agg (average is used in our experiments), our approach has a significant improvement over NF [59] (4.8mm) and MHFormer [33] (3.1mm), respectively. However, the increase of H (1→20) does not significantly improve the performance of the proposed D3DP method under P-Best (40.0→39.5mm). This result prompts us to propose the new setting J-best, under which a noticeable improvement (40.0→35.4mm) can be observed. It is because the same hypothesis may perform differently depending on which joint is being considered. This observation motivates us to propose J-Agg (using our JPMA) to exploit the joint-level differences between hypotheses. Compared to P-Agg, the performance under J-Agg is boosted (39.9→39.5mm). This performance is the same as P-Best (39.5mm). More results in terms of P-MPJPE or using 2D ground truth as inputs are provided in Appendix E.

Method		PCK↑	AUC↑	MPJPE↓
U-CDGCN [21] (N=96)▽	MM'21	97.9	69.5	42.5
P-STMO [47] (N=81)▽	ECCV'22	97.9	75.8	32.2
TCN [42] (N=81)	CVPR'19	86.0	51.9	84.0
U-GCN [58] (N=96)	ECCV'20	86.9	62.1	68.1
Anatomy [9] (N=81)	TCSVT'21	87.9	54.0	78.8
PoseFormer [65] (N=9)	ICCV'21	88.6	56.4	77.1
MHFormer [33] (N=9, H=3, P-Agg)	CVPR'22	93.8	63.3	58.0
MixSTE [64] (N=27)	CVPR'22	94.4	66.5	54.9
MixSTE [64] (N=243)‡	CVPR'22	96.9	75.8	35.4
D3DP (N=243, H=1, K=1)		97.7	77.8	30.2
D3DP (N=243, H=20, K=20, P-Agg)		97.7	78.0	30.0
D3DP (N=243, H=20, K=20, J-Agg)		97.7	78.2	29.7
D3DP (N=243, H=20, K=20, P-Best♯)		97.8	78.2	29.8
D3DP (N=243, H=20, K=20, J-Best♯)		98.0	79.1	28.1

Table 2: Results on MPI-INF-3DHP under three evaluation metrics using ground truth 2D keypoints as inputs. (▽) - They train and evaluate on 3D poses scaled to the height of the universal skeleton used by H36M, while we use the ground truth 3D poses.

MPI-INF-3DHP. Table 2 shows comparisons between our method and others on 3DHP dataset. D3DP reduces the MPJPE by 5.2mm on top of MixSTE [64] in the single-hypothesis case (H=1). When H is increased to 20, the performance under J-Agg using JPMA is similar to P-Best and superior to P-Agg. Our method performs best when evaluated under J-Best. These results exhibit the same pattern as those on Human3.6M, which demonstrates the strong generalization capability of our method.

4.4. Qualitative Results

Fig. 5 shows the qualitative results of D3DP on three datasets with H, K set to 5. For the cases of 2D keypoint misdetection (top left: right knee), occlusion (top left: lower body, bottom right: right arm), and depth ambiguity (bottom left: upper body), our method generates pose hypotheses with high variance to reflect the uncertainty in 3D pose estimation. For simple poses (top/bottom middle), D3DP is inclined to generate a deterministic prediction. For rare poses (top right), although the predicted 3D hypotheses are not accurate, they are anatomically plausible and diverse in

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

Diffusion	JPMA	H	Setting	MPJPE↓
		1	N/A	41.0
✓		1	N/A	40.0
✓		20	P-Agg	39.9
✓	✓	20	J-Agg	39.5

(a) Effectiveness of components.

Level	Method	MPJPE↓
pose	average	39.9
pose	MLPs	42.5
joint	MLPs	41.6
pose	reproj.	39.7
joint	reproj.	39.5

(b) Multi-hypothesis aggregation.

Table 3: Ablation experiments on Human3.6M. H=20, K=10.

Method	H	K	MPJPE↓	Params(M)	Train. MACs(G)	Infer. MACs(G)	Infer. FPS
PoseFormer [65]	1	N/A	44.3	9.5	0.81	1.62	1952
MHFormer [33]	3	N/A	43.1	24.7	4.81	9.62	598
P-STMO [47]	1	N/A	42.8	6.7	0.87	1.74	3040
MixSTE [64]	1	N/A	41.0	33.6	0.57	1.14	4547
D3DP	1	1	40.0	34.6	0.57	1.14	4364
D3DP	5	5	39.7	34.6	0.57	28.5	204
D3DP	20	10	39.5	34.6	0.57	228.2	29

Table 4: Accuracy&speed. MACs are averaged over each frame. The proposed method uses the J-Agg setting.

the arm region to approximate the correct result. More results are provided in Appendix F.

4.5. Ablation Study

We conduct ablation experiments on Human3.6M to validate the impact of each design in our method. More results are provided in Appendix D.

Effectiveness of each component. As shown in Table 3a, we begin with MixSTE [64] as the backbone (first row), and then combine it with the diffusion models, resulting in our D3DP (second row), which achieves 1mm improvement in the single-hypothesis case. This result reveals that a simple combination of diffusion models and existing deterministic pose estimators brings a large performance gain. Besides, the network is equipped with the ability to generate multiple pose hypotheses, which fits the property of depth uncertainty in 3D pose estimation. Under P-Agg setting, MPJPE slightly decreases (40.0→39.9mm) when H is raised to 20. The error is further reduced to 39.5mm by using our JPMA (under J-Agg) method. This result shows that multi-hypothesis aggregation at the joint level, which allows fine-grained hypothesis selection, is superior to that at the pose level.

Multi-hypothesis aggregation. We compare pose-level and joint-level multi-hypothesis aggregation methods in Table 3b, including 1) averaging over all hypotheses; 2) using a 4-layer MLPs to generate scores for each pose/joint, which are used for a weighted average of hypotheses; 3) using our reprojection-based method at the pose level (PPMA, 4th row) and joint level (JPMA, 5th row). For fairness, we implement the same approach (MLPs/reprojection-based) at both levels. The results show that joint-level approaches are superior to pose-level approaches. Besides, PPMA outperforms both MLPs and average at the pose level and

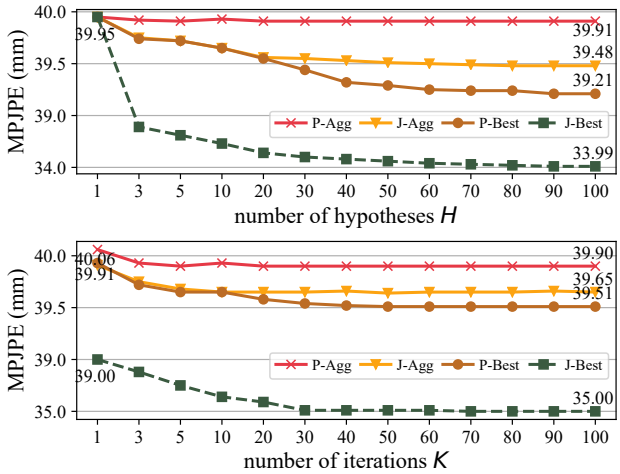

Figure 6: Ablation experiments on the number of hypotheses and iterations H, K. Top: K=10. Bottom: H=10.

JPMA outperforms MLPs at the joint level, which shows that the 2D prior is effective in multi-hypothesis aggregation. Thus, our JPMA, which conducts aggregation at the joint level using 2D priors, achieves the best performance.

Accuracy and speed. Table 4 shows the accuracy and inference speed. Experiments are conducted on a single NVIDIA RTX 3080 Ti GPU. For previous methods (first four rows), MACs of inference are twice as high as those of training, due to the data augmentation applied. D3DP (5th row) makes a few modifications (add additional inputs and timestep embedding) to MixSTE (our backbone, 4th row), and adds a parameter-free diffusion process. This results in a small increase in the number of parameters, essentially unchanged MACs, a slight decrease in FPS, and a 1mm performance improvement. Moreover, we can adjust H, K to trade computational budgets for performance during inference. When H, K increases, MACs and FPS roughly show a linear increase and decrease, respectively.

Number of hypotheses and iterations. Fig. 6 (Top) illustrates how performance fluctuates with the number of hypotheses under the four settings. They all start with the same performance (H=1) and then gradually diverge. 1) P-Best and J-Best: The error under J-Best is much lower than that under P-Best, which indicates that joint-level aggregation (J-Agg) has a higher upper-bound performance than pose-level aggregation (P-Agg). 2) P-Agg and J-Agg: Average and JPMA are used as the pose-level (P-Agg) and joint-level (J-Agg) methods, respectively. MPJPE under P-Agg stays around 39.9mm. It is because as H increases, the mean of all hypotheses remains roughly the same, while the variance rises (Appendix D.3). This observation indicates that the average operation counteracts the advantage of increased diversity of hypotheses, leading to no improve-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

ment in performance. The error under J-Agg decreases as H rises, which demonstrates that J-Agg makes better use of the diversity than P-Agg to achieve performance gains when H increases. The performance of J-Agg is very close to the upper-bound performance of P-Agg (*i.e.*, P-Best) when $H < 20$. J-Agg is practical, but P-Best is not because the latter is agnostic in real-world applications, which highlights the advantages of J-Agg. Besides, there is still a performance gap between J-Agg (39.48mm) and J-Best (33.99mm), which shows that JPMA serves as a baseline method with considerable room for improvement.

Fig. 6 (Bottom) illustrates how performance fluctuates with the number of iterations. The general trend of the performance curves for the four settings is similar to that of the above experiment. Performance under P-Agg does not vary substantially with K. J-Agg and P-Best diverge after K=10. J-Best outperforms the other three settings by a wide margin. The performance of all settings saturates at K=30.

5. Conclusions

This paper presents D3DP with JPMA, a diffusion-based probabilistic method with joint-wise reprojection-based multi-hypothesis aggregation, for 3D human pose estimation. D3DP, which is compatible and customizable, generates multiple pose hypotheses. It progressively adds noise to clean 3D poses and reverses this process to reconstruct the original signals. To aggregate the generated pose hypotheses into a single 3D prediction, JPMA is proposed, which conducts joint-level aggregation based on reprojection errors. Compared with existing pose-level aggregation methods, JPMA exploits the merits of each joint and thus achieves exceptional performance. Experimental results on two benchmarks show that our method surpasses the state-of-the-art probabilistic and deterministic approaches.

Acknowledgment

This work was supported in part by the NSFC under grant No. 62025101, 61931014 and U21B2012, and in part by the Science and Technology Plan Project in Basic Public Welfare class of Shaoxing city(No.2022A11002), which are gratefully acknowledged.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 2

[3] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 2

[4] Christopher M Bishop. Mixture density networks. 1994. 3

[5] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4186, 2022. 2

[6] 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*, pages 7291–7299, 2017. 1

[7] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith Mv, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019. 3

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

[9] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 6, 7

[10] 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

[11] 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*, pages 7103–7112, 2018. 1, 6

[12] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 723–732, 2019. 3

[13] Jeongjun Choi, Dongseok Shim, and H Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *arXiv preprint arXiv:2212.02796*, 2022. 3

[14] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2271, 2019. 1

[15] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 2

[16] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 4

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 2

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

[21] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional directed graph convolution for 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 602–611, 2021. 6, 7

[22] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022. 2

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5

[24] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 805–814, 2017. 1, 4

[25] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, pages 10362–10383. PMLR, 2022. 2

[26] Sungwon Kim, Heeseung Kim, and Sungroh Yoon. Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *arXiv preprint arXiv:2205.15370*, 2022. 2

[27] Jogendra Nath Kundu, Siddharth Seth, MV Rahul, Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11312–11319, 2020. 3

[28] Mun Wai Lee and Isaac Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004. 3

[29] Alon Levkovitch, Eliya Nachmani, and Lior Wolf. Zero-shot voice conditioning for denoising diffusion tts models. *arXiv preprint arXiv:2206.02246*, 2022. 2

[30] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 1, 2, 3, 4, 6

[31] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3d human pose hypotheses. *BMVC*, 2020. 2, 3, 5, 6, 7

[32] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 2022. 6

[33] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 2, 3, 4, 5, 6, 7, 8

[34] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 1

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 1

[37] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 5

[38] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[39] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4484. PMLR, 2020. 2

[40] Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021. 1, 2, 3, 4, 5, 6, 7

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[42] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 1, 3, 5, 6, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.

[43] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2, 3, 4

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[45] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[46] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[47] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022. 1, 5, 6, 7, 8

[48] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3446–3454, 2021. 6

[49] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2325–2334, 2019. 1, 2, 3, 4, 6, 7

[50] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenya, Carme Torras, and Francesc Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2673–2680. IEEE, 2012. 3

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[52] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 2, 3, 5

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 5, 6

[55] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022. 2

[56] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 6

[57] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7782–7791, 2019. 3

[58] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 7

[59] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11199–11208, 2021. 1, 2, 3, 5, 6, 7

[60] Qi Yan, Zhengyang Liang, Yang Song, Renjie Liao, and Lele Wang. Swingnn: Rethinking permutation invariance in diffusion models for graph generation. *arXiv preprint arXiv:2307.01646*, 2023. 2

[61] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020. 1, 6

[62] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13116–13125, 2022. 6

[63] Jinlu Zhang, Yujin Chen, and Zhigang Tu. Uncertainty-aware 3d human pose estimation from monocular video. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5102–5113, 2022. 6

[64] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022. 1, 5, 6, 7, 8

[65] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *arXiv preprint arXiv:2103.10455*, 2021. 1, 6, 7, 8

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:28:54 UTC from IEEE Xplore. Restrictions apply.