SnapPose3D: Diffusion-Based Single-Frame 2D-to-3D Lifting of Human Poses

Alessandro Simoni[1], Riccardo Catalini[1], Davide Di Nucci[1], Guido Borghi[1], Davide Davoli[2*], Lorenzo Garattoni[2], Gianpiero Francesca[2], Yuki Kawana[3], and Roberto Vezzani[1]

[1] University of Modena and Reggio Emilia (UniMoRe), Italy
[2] Toyota Motor Europe (TME), Brussels, Belgium (*providing contracted services)
[3] Woven by Toyota, Japan

Abstract. Depth ambiguity and joint uncertainty are the two main obstacles in obtaining accurate human pose predictions by 2D-to-3D lifting methods proposed in the literature. In particular, these issues are caused by 2D joint locations that can be mapped to multiple 3D positions, inducing multiple possible final poses. Following these considerations, we propose leveraging diffusion-based models' generation capability to predict multiple hypotheses and aggregate them in a final accurate pose. Therefore, we introduce SnapPose3D, a pose-lifting framework trained deterministically to denoise 3D poses conditioned on both visual context and 2D pose features. SnapPose3D adopts a probabilistic approach during inference, generating multiple hypotheses through random sampling from a unit Gaussian distribution. Unlike most previous methods that address pose ambiguity by processing temporal sequences, SnapPose3D uses single frames as input, avoiding tracking and limiting computational cost, data acquisition complexity, and the need for online, real-time applications. We extensively evaluate SnapPose3D on well-known benchmarks for the 3D human pose estimation task showing its ability to generate and aggregate accurate hypotheses that lead to state-of-the-art results.

Keywords: 3D Human Pose Estimation, Diffusion Model

1 Introduction

The estimation of human poses from single images has been a topic of significant interest and importance for several years in the computer vision community, owing to its wide range of applications, including human-robot interaction [41, 47], worker safety [10, 32], and social behavior analysis [1, 27]. However, while estimating the 2D joint positions on images is a well-defined problem and numerous methods already achieve extreme accuracy [3, 45, 54] on well-established benchmark datasets [23, 25, 29], the estimation of the 3D joint positions still poses significant challenges.

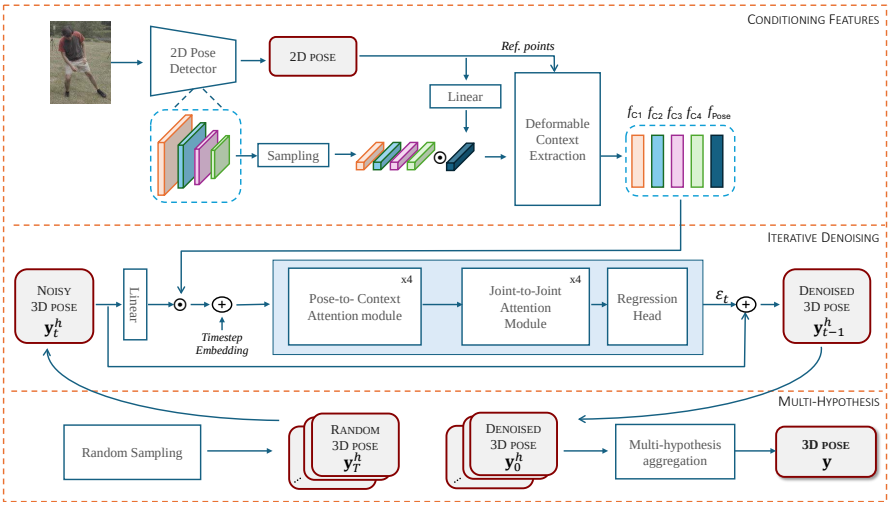

Fig. 1: Overview of SnapPose3D framework composed of three steps: (i) extraction of the conditioning features, *i.e.* visual features computed started from the initial 2D pose and the context near the subject; (ii) then, a transformer-based architecture iteratively removes the noise ϵ_t from each input pose $\mathbf{y}_t^h$; (iii) finally, the final pose is predicted by aggregating H multiple denoised hypotheses.

The main issues are related to the depth ambiguity and joint location uncertainty, due to the inherent loss of three-dimensional information when projecting to monocular 2D. Besides, difficulties in collecting 3D data and annotations in realistic settings further increase the challenges [50]. Despite these challenges, 3D Human Pose Estimation (3D HPE) [36, 40] from single images is an important task and attracts significant interest in the computer vision community.

In this context, we introduce **SnapPose3D**, a diffusion-based framework that accurately estimates 3D human poses directly from a single input image.

Following a common trend in the literature [7, 34, 60], the 2D pose estimation procedure represents the starting point, due to the excellent performance of currently available methods [4, 5, 45]. Subsequently, the initial 2D pose is lifted to a set of multiple and plausible 3D poses through a conditioned diffusion-based mechanism [43]; these poses are then finally aggregated in a single 3D pose.

We show that the advantages of generating multiple hypotheses are twofold: firstly, their aggregation is a key element in solving the depth ambiguity since it allows for discarding outliers or less plausible predictions. Secondly, the analysis of the distribution of the hypotheses enables the inference of the degree of confidence in the predictions. From a technical perspective, SnapPose3D is able to generate different hypotheses through a transformer-based architecture [48], leveraging the knowledge of conditioning features, *i.e.* the initial 2D human pose and multi-resolution visual features extracted from the contest around the

subject. Additionally, SnapPose3D is based on a single frame as input: we deliberately refrain from incorporating temporal information as done by most of the previous works [6, 11, 37], then reducing the complexity of data acquisition and enabling online applications. Indeed, temporal data integration typically involves estimating a human pose across consecutive frames using, for instance, a sliding window or more complex techniques. However, such a processing approach is feasible only in offline scenarios where subsequent frames are readily available or where a degree of latency in result delivery is acceptable. Moreover, temporal tracking becomes imperative in multi-person scenarios. Conversely, estimating the 3D pose from a single frame can be done before tracking individuals.

We prove the efficacy of SnapPose3D through the evaluation of two well-known benchmark datasets for 3D human pose estimation, namely Human3.6M [16] (including its hard subset H36MA [52]) and MPI-INF-3DHP [29]. The experimental results show that our method outperforms the performance of sota competitors. In particular, we showcase a robust correlation between the reliability of the estimated pose and the similarity among the generated hypotheses. Finally, a qualitative evaluation of in-the-wild videos from 3DPW [49] is reported.

2 Related work

2.1 3D Human Pose Estimation (3D HPE)

Most methods for 3D HPE are based on a preliminary 2D pose estimation phase followed by a lifting step that recovers the 3D original position starting from its 2D projection. Thanks to the high reliability of state-of-the-art 2D pose estimators [5, 45], the retrieval of the missing third coordinate is the most challenging part of 3D HPE. Without any additional information, the 2D-to-3D lifting is almost an ill-posed problem. Thus, temporal information [24, 34, 60], available if processing a sequence of 2D postures at a time, or visual features of the person and the corresponding context are required to inject more constraints.

The main limitations of the first class of methods are related to the latency generated by estimating the output of the central frame within a window and the need for tracking in case of multi-person scenarios. On the contrary, the second class of methods requires more computational load to reprocess the appearance features, as already done in the preliminary 2D pose estimation step.

As confirmed in [9, 39, 42], the variability of human postures does not allow the application of sufficient constraints to back-project 2D joints into the original 3D space. However, multiple hypotheses could be formulated to simultaneously satisfy the alignment to the 2D pose and additional conditions, such as anatomical constraints. For instance, Jahangiri et al. [17] apply anatomical constraints such as joint angle and bone length to limit the number of hypotheses. Li et al. [24] proposes a multi-hypothesis Transformer to predict a 3D pose of a central frame, but it requires a long input sequence of 2D keypoints as additional temporal information. Li et al. [22] combines a mixture density network [2] with the inverse problem of monocular 3D-HPE to generate plausible 3D poses.

Oikarinen et al. [30] improves the mixture density network-based method [22] with the strength of semantic graph neural networks [57].

Conditional variational autoencoders have been exploited by Sharma et al. [38] to sample 3D-pose candidates that are aggregated with a specific deep network to obtain the final prediction. Wehrbein et al. [52] model the posterior distribution of 3D poses with normalizing flow by explicitly incorporating the uncertainty provided by the detector together with the 2D keypoint positions.

2.2 Diffusion Models for 3D HPE

Following the impressive results obtained by Denoising Diffusion Probabilistic Models (DDPM) [13] in image synthesis [8, 35, 44], the same approach has been applied to the 3D-HPE problem. In the HPE case, the human pose is considered a non-equilibrium system that diffuses to a completely random set of joint coordinates. Given this assumption, the DiffPose framework [11] considers the 3D pose estimation as a reverse diffusion process, where a 3D pose distribution with high uncertainty and indeterminacy is progressively transformed toward a 3D pose with low uncertainty. To guide the diffusion model, DiffPose leverages the spatial-temporal context extracted from a 2D pose sequence. However, a single-person video or a tracking algorithm is required.

A key advantage of DDPMs for pose estimation is the natural generation of multiple hypotheses. A concurrent diffusion-based method, also named DiffPose [14], conditions the process on 2D keypoint heatmaps from a single frame, limiting its applicability to heatmap-based methods. D3DP [37] exploits this probabilistic nature to enhance performance through hypothesis aggregation.

A joint-wise reprojection-based multi-hypothesis aggregation (JPMA) method is proposed for limiting the influence of outliers when the intrinsic calibration parameters of the cameras are available. As the human pose has tight semantic connectivity between adjacent joints, Choi et al. [6] adopt a graph convolutional neural network (GCN) [21] as a denoiser architecture to explicitly learn correlations between joints.

3 SnapPose3D

Given an image $I^{H\times W\times 3}$ as input, the goal of SnapPose is to predict the 3D human pose $\mathbf{y} \in \mathbb{R}^{J\times 3}$ where J is the number of skeleton's joints.

As visually summarized in Figure 1, the diffusion process is conditioned with features embedding the visual context near the subject – which is usually discarded by temporal pose-lifting architectures – and the 2D pose. Both of them are computed on the single input image.

3.1 Preliminary

Diffusion models [43] represent probabilistic generative approaches that learn to synthesize data through a parametrized Markov chain with variational inference. In particular, a diffusion model is composed of two recurrent processes:

(i) *forward diffusion* that creates noisy samples starting from a target data x_0, and (ii) *reverse diffusion* that progressively denoises a random Gaussian noise to get the sample x_0. In the following, we follow the formulation of Denoising Diffusion Probabilistic Models (DDPM) [13]. To learn the reverse diffusion process, a set of intermediate noisy samples is needed to bridge the gap between the target sample and the Gaussian noise. The forward process is applied where the posterior distribution $q(x_{1:T} \mid x_0)$ is computed by gradually adding infinitesimal Gaussian noise ϵ with a variance $\beta_t \in [0, 1]$ for each timestep $t \in [0, T]$:

$$q(x_{1:T} \mid x_0) := \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \quad q(x_t \mid x_{t-1}) := \mathcal{N}_{pdf}(x_t \mid \sqrt{1-\beta_t}\, x_{t-1}\ ,\ \beta_t \mathbf{I}) \tag{1}$$

where $\mathcal{N}_{pdf}$ is the likelihood of sampling x_t given the conditioning parameters, T is the total number of diffusion steps, and $\mathbf{I}$ denotes the identity matrix.

On the other hand, through the reverse process, the goal is to estimate the posterior $q(x_{t-1} \mid x_t)$ and thus find the target data x_0 or its distribution. Given $\alpha := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{k=1}^{t} \alpha_k$, we can apply the Bayes theorem and formulate $q(x_{t-1} \mid x_t)$ as follows:

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \qquad \tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{2}$$

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}_{pdf}(\tilde{\mu}_t(x_t, t), \tilde{\beta}_t \mathbf{I}) \tag{3}$$

Since the above formulation needs to know x_0 in advance, diffusion models leverage neural networks to estimate the value of the posterior $p_\theta(x_{t-1} \mid x_t) \simeq q(x_{t-1} \mid x_t, x_0)$. Thus, a diffusion model aims to approximate the reverse process through a neural network learning the infinitesimal noise $\epsilon(x_t, t)$ between two consecutive timesteps. The objective loss function can be defined as $\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}[\| \epsilon - \epsilon_\theta(x_t, t) \|^2]$.

3.2 Conditioning Features extraction

The conditioning of the reverse denoising is obtained with a set of features obtained by merging together appearance and 2D pose information. Firstly, HRNet-32 [45] is used as a multi-resolution feature extractor to generate $f'_C = f'_{C1} \odot f'_{C2} \odot f'_{C3} \odot f'_{C4}$, where $\odot$ is the concatenation operator and $f'_{C1} \dots f'_{C4}$ are hierarchical feature vectors (from $L = 4$ levels) converted to a common size through sampling. In addition, HRNet-32 also generates the 2D pose, projected by a linear layer to f'_P and concatenated to the visual features. Eventually, a pretrained Deformable Context Extraction module [58] generates the final multichannel descriptor $F \in \mathbb{R}^{(L+1)\times J\times d}$, where $L = 4$ is the number of feature levels, increased by one with the pose channel, J is the number of joints and d is the embedding size ($d = 128$ in our experiments).

3.3 Iterative Denoising procedure

The denoising network is a transformer-based network that takes as input the embedding previously defined and concatenated to a projection of the current 3D noisy pose. A timestep embedding f_t is added to the concatenated features to make the model aware of the current diffusion step. Formally, the noisy pose $\mathbf{y}_t \in \mathbb{R}^{J\times 3}$ is linearly projected into a higher-dimensional embedding of dimension $d = 128$. The resulting input embedding is processed by two transformer-based modules that perform self-attention between pose and context and between each joint, respectively. Lastly, a regression head outputs the predicted noise $\hat{\epsilon}$ that was initially added to the ground truth pose $\mathbf{y}_0$.

The Pose-to-Context Attention Module considers each channel as a token of size $d \times J$, in which d is the size of the embedding. The Joint-to-Joint Attention Module considers each joint as a token of size $c \times d$, where c is the total number of channels. The first module enriches each joint with contextual information, while the second one allows data sharing between the different joints.

We note that to apply the diffusion process to the 3D human pose estimation task, previous works [6, 11, 37] took inspiration from the 2D-to-3D lifting literature, adapting state-of-the-art architectures as a denoiser function. In particular, existing lifting approaches [24, 60] predict the 3D pose of the central frame from a temporal sequence of 2D poses. However, this technique defines a *non-causal* system requiring access also to future poses, that are not available in online real-world applications. In addition, a tracking system is required in multi-person scenarios, and the output is given with a non-negligible latency.

3.4 Training and Inference

During training, a timestep $t \in [0, T]$ is randomly selected for each sample in the batch. Subsequently, the forward diffusion process is applied to the ground truth 3D pose $\mathbf{y}_0$ to obtain the noisy pose $\mathbf{y}_t$. Following the definition in Equation 1, this operation can be formulated as $q(\mathbf{y}_t \mid \mathbf{y}_0) := \prod_t q(\mathbf{y}_t \mid \mathbf{y}_{t-1})$. The resulting noisy pose $\mathbf{y}_t$ is given as input to our denoiser network $\mathcal{D}$ conditioned on context features f_C, 2D pose features f_P and timestep embedding f_t to predict the added noise $\hat{\epsilon}$: $\hat{\epsilon} = \mathcal{D}(\mathbf{y}_t, f_C, f_P, f_t)$. The objective loss of the proposed method is an MSE loss computed on the noise as follows: $\mathcal{L} = \mathbb{E}_{\mathbf{y}_0, t, \epsilon\sim\mathcal{N}(0,\mathbf{I})}[\| \epsilon - \hat{\epsilon} \|^2]$

Since the corrupted pose $\mathbf{y}_t$ with $t \to T$ can be approximated by a Gaussian distribution, the reverse process can start from H initial hypotheses $\mathbf{y}_T^h, h \in [1, H]$, each one obtained by sampling random noise from a unit Gaussian. These pure-noisy hypotheses are given as input to the denoiser $\mathcal{D}$ that, conditioned on the visual features extracted from the current frame, outputs the quantity of noise $\hat{\epsilon}_t$ which should be removed to get the 3D pose predictions $\mathbf{y}_{t-1}^h$ at the next timestep. During inference, we use an optimized version of DDPM denoising strategy, namely DDIM [44]. For each hypothesis in parallel, the predicted noise $\hat{\epsilon}$ is removed from the corrupted pose as follows $\mathbf{y}_{t-1}^h = \mathbf{y}_t^h - \hat{\epsilon}_t$. Starting with $t = T$, this procedure is iteratively repeated K times (K=20 in our experiments) so that the timestep for each iteration is defined as $t = T \cdot (1 - k)/\mathrm{K}, k \in [0, \mathrm{K})$.

3.5 Multiple Hypothesis Aggregation

As mentioned, the advantage of using a diffusion-based approach like SnapPose3D is that it combines the simplicity of deterministic training with the flexibility of probabilistic inference. Indeed, during training, SnapPose3D produces a single prediction while learning to handle Gaussian noise added to each sample. However, during inference, it generates multiple hypotheses by sampling from a Gaussian distribution, allowing an arbitrary number H of predictions.

Following this approach, it is possible to output multiple plausible poses conditioned on the same input frame. If the intrinsic camera parameters are given, the re-projection error of the 3D poses when compared with the initial 2D poses could be exploited [37] as a selection schema, even if we prefer to ignore them, even when available, as in the most general case. A single hypothesis can be randomly selected as the final pose estimation among the available H ones (called **R**andom selection). The output pose can also be obtained by selecting the best hypothesis, *i.e.* the one closest to the ground truth (**B**est selection). This approach is also referred to as "Supervision from an Oracle" [38], but it is important to note that it requires the knowledge of the ground truth.

However, generating multiple hypotheses not only increases the probability of obtaining at least one good solution but also produces additional information that can be exploited through proper hypothesis aggregation. Given H hypotheses $\mathbf{y}_0^h$, a unique solution $\mathbf{y}_0$ must be produced as output. Averaging (**A**) is the standard aggregation approach, though it is sensitive to outliers. To address this, we also employ the Median operator (**M**), which selects the central element after independently sorting the $3 \times J$ vectors, each containing the H hypotheses for each joint coordinate (X, Y, Z).

4 Datasets and metrics

4.1 Datasets

Human3.6M [16] is a large-scale dataset containing 3.6 million images with annotated 3D human poses with a skeleton of 17 joints. The dataset has been captured using four cameras from different views in an indoor environment. A total of 11 subjects perform 15 daily activities. To allow a fair comparison with the state-of-the-art, we train on subjects (S1, S5, S6, S7, S8), while we test on subjects (S9, S11). Furthermore, to underline the advantage of having a multi-hypothesis system and following the recent DiffPose [14] work, we also tested SnapPose3D on the **H36MA** subset of Human3.6M defined by [52]. H36MA contains 6.4% of the test set, which includes highly ambiguous cases.

MPI-INF-3DHP this dataset [29] has been captured with 14 cameras placed in both indoor and outdoor environments. More than 1.3 million frames have been captured and annotated using a skeleton model with 17 joints. The frames contain 8 actors performing 8 activities such as walking, sitting, sport, and dynamic actions. Following the official train/test split, we train our method using 8 camera views of the training set and evaluate the valid frames of the test set.

Table 1: Results in terms of MPJPE on Human3.6M using CPN [5] 2D pose detections as input. Competitors are subdivided into three groups from top to bottom: (i) baselines (ii) diffusion approaches providing a single prediction through aggregation, (iii) diffusion approaches that select one of the hypotheses as output. For each part, best results are in bold, second ones are underlined. Number of hypotheses H and aggregation (**A**verage, **M**edian) or selection (**R**andom, **B**est) method are also reported.

Protocol 1 (**MPJPE**) (mm)	Dir	Dis	Eat	Gre	Phn	Pht	Pos	Pur	Sit	SitD	Smk	Wait	WlkD	Walk	WlkT	**Avg**
Baseline methods																
Pavlakos [33]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Martinez [28]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhao [57]	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
Sun [46]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Yang [55]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Hossain [15]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Liu [26]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	**46.0**	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Xu [53]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	**59.7**	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Zhao [59]	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	**39.7**	43.1	51.8
Zhao [58]	**37.7**	**42.0**	**39.0**	**42.7**	**43.5**	**37.6**	**41.1**	51.4	61.7	**43.4**	**49.4**	**43.0**	**33.9**	45.2	**35.0**	**43.4**
Diffusion-based methods Aggregation of multiple hypotheses																
(H=10,A) Diffupose [6]	43.4	50.7	45.4	50.2	49.6	53.4	48.6	**45.0**	56.9	70.7	47.8	48.2	51.3	43.1	43.4	49.4
(H=5,A) Diffpose [11]	42.8	49.1	45.2	48.7	52.1	63.5	46.3	45.2	58.6	66.3	50.4	47.6	52.0	**37.6**	40.2	49.7
(H=20, A) *SnapPose3D*	**36.4**	41.8	**39.6**	42.1	42.7	**36.7**	**43.0**	53.5	62.0	**42.1**	51.3	40.6	**32.0**	45.3	35.3	43.2
(H=20, M) *SnapPose3D*	43.2	**41.3**	41.4	**39.0**	**40.5**	44.8	43.9	47.5	**46.2**	50.5	**43.2**	**40.0**	41.5	40.5	**34.5**	**42.8**
Diffusion-based methods One hypothesis selected																
(H=200, B) DiffPose [14]	38.1	43.1	**35.3**	43.1	46.6	48.2	39.0	**37.6**	51.9	59.3	41.7	47.6	45.4	37.4	36.0	43.3
(H=20, R) *SnapPose3D*	38.7	44.4	41.6	44.6	45.0	38.8	45.6	55.6	65.3	44.4	54.1	43.1	34.3	48.2	37.7	45.6
(H=20, B) *SnapPose3D*	**37.5**	**35.5**	35.7	**33.4**	**34.9**	**38.6**	**37.8**	41.1	**40.0**	**43.7**	**37.4**	**34.5**	**35.8**	**34.8**	**29.2**	**36.9**

4.2 Metrics

The Mean Per Joint Position Error (MPJPE) [18] calculates the average Euclidean distance between estimated joint 3D coordinates and their ground truth in millimeters is used. Moreover, on Human3.6M we also use the aligned version of the MPJPE (usually referred to P-MPJPE or reconstruction error [19]), which is MPJPE after rigid alignment of the prediction with ground truth via Procrustes Analysis [12]. On MPI-INF-3DHP the performance is assessed also in terms of the Percentage of Correct Keypoints (PCK) within a 150mm range and Area Under Curve (AUC) metrics.

5 Experimental Results

Taking inspiration from previous literature methods [37], we design the SnapPose3D framework, adapting a transformer-based deterministic approach [58] to the diffusion model. In particular, the first attention module is a 4-layer multi-head attention block that learns relations between context and pose features. The second attention module is similarly implemented in terms of learnable functions,

Table 2: Results in terms of P-MPJPE on Human3.6M using CPN [5] 2D pose detections as input. Further details are reported in the caption of Table 1.

Protocol 2 (**P-MPJPE**) (mm)	Dir	Dis	Eat	Gre	Phn	Pht	Pos	Pur	Sit	SitD	Smk	Wait	WlkD	Walk	WlkT	**Avg**
Baseline methods																
Sun [46]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Martinez [28]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Hossain [15]	36.9	37.9	42.8	40.3	46.8	46.7	37.7	36.5	48.9	52.6	45.6	39.6	43.5	35.2	38.5	42.0
Pavlakos [33]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Liu [26]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	**31.7**	36.2	41.2
Yang [55]	**26.9**	**30.9**	36.3	39.9	43.9	47.4	**28.8**	**29.4**	**36.9**	58.4	41.5	**30.5**	29.5	42.5	32.2	37.7
Zhao [58]	32.3	34.6	**33.1**	**35.1**	**35.2**	**30.1**	32.1	42.6	48.8	**36.0**	**39.1**	33.1	**27.5**	37.1	**29.3**	**35.4**
Diffusion-based methods	**Aggregation of multiple hypotheses**															
(H=10,A) Diffupose [6]	35.9	40.3	36.7	41.4	39.8	43.4	37.1	35.5	46.2	59.7	39.9	38.0	41.9	32.9	34.2	39.9
(H=5, A) Diffpose [11]	33.9	38.2	36.0	39.2	40.2	46.5	35.8	**34.8**	48.0	52.5	41.2	36.5	40.9	**30.3**	33.8	39.2
(H=20, A) *SnapPose3D*	**31.5**	34.7	**32.7**	35.1	34.2	**29.9**	**33.1**	42.8	47.5	**34.4**	38.6	31.6	**26.5**	36.7	29.7	34.8
(H=20, M) *SnapPose3D*	35.7	**34.3**	34.0	**32.5**	**32.9**	36.7	35.6	38.2	**37.1**	38.4	**34.3**	**31.5**	32.8	32.4	**29.0**	**34.5**
Diffusion-based methods	**One hypothesis selected**															
(H=200, B) DiffPose [14]	**28.1**	31.5	**28.0**	30.8	33.6	35.3	**28.5**	**27.6**	40.8	44.6	31.8	32.1	32.6	28.1	26.8	32.0
(H=20, R) *SnapPose3D*	34.4	37.7	35.1	38.1	36.9	32.4	36.3	45.5	51.7	37.2	42.0	34.6	29.2	40.1	32.6	37.8
(H=20, B) *SnapPose3D*	30.8	**29.2**	29.0	**27.8**	**28.2**	**31.6**	30.6	33.0	**32.1**	**33.3**	**29.6**	**27.2**	**28.3**	**27.9**	**24.5**	**29.7**

but it learns correlations between the skeleton joints. Finally, a regression head projects the feature space from d to 3 dimensions.

We train SnapPose3D with a batch size of 128 for 50 epochs on both datasets. We use Adam optimizer [20] with a starting learning rate of $6e^{-4}$ using a linear decay with factor 0.993. We apply random horizontal flipping as data augmentation for both images and skeletons. We implement SnapPose3D using the PyTorch [31] framework and the *Diffusers* [1] library for the diffusion-based technique. Moreover, to guarantee the full reproducibility of our results, the code and the training/testing hyperparameters will be publicly released.

In this section, we focus on quantitative results, and refer the reader to the supplementary material for qualitative evaluations.

[1] https://huggingface.co/docs/diffusers/index

5.1 Comparison with state-of-the-art

To make the comparison with state-of-the-art techniques more fair, we directly compared SnapPose3D with single-frame approaches, able to work without the need for the temporal tracking of people. Thus, we focus on single-frame approaches, dividing diffusion-based methods from the others for clarity.

Table 3: Results on the hard subset H36MA in terms of MPJPE and P-MPJPE. Number of hypotheses H and aggregation (**A**verage, **M**edian) or selection (**R**andom, **B**est) method are also reported (see Sect. 3.5).

Method	MPJPE (mm)	P-MPJPE (mm)
Li et al. [22]	81.1	66.0
Sharma et al. [38]	78.3	61.1
Wehrbein et al. [52]	71.0	54.2
(H=20, A) *SnapPose3D*	<u>46.2</u>	<u>36.5</u>
(H=20, M) *SnapPose3D*	**45.8**	**36.2**
(H=200, B) DiffPose [14]	63.1	46.7
(H=1, R) *SnapPose3D*	<u>48.8</u>	<u>39.8</u>
(H=20, B) *SnapPose3D*	**39.6**	**31.6**

Evaluation on Human3.60M. In Table 1 and Table 2, we report the results in terms of MPJPE (Protocol 1) and P-MPJPE (Protocol 2) on the test set of Human3.6M. Following previous works that use CPN [5] to get the 2D joints as input, we exploit the same network as the first step of our method. We tested SnapPose3D using all the selection and aggregation approaches described in Section 3.5. As reported, SnapPose3D surpasses its direct competitors [6, 11, 14] and achieves state-of-the-art performance across various actions and in the overall average score. Notably, the use of the median aggregation (**M** aggregation) yields even better results with respect to the average one (**A** aggregation), with a final average error of 42.8mm using Protocol 1 and 34.5mm using Protocol 2.

Noticeably, even when selecting a random hypothesis among the 20 generated ones (H=20, **R** selection), SnapPose3D achieves high accuracy, only slightly inferior to that of the competitors. This result underscores the robustness and effectiveness of the proposed method, which generates plausible candidate poses.

As expected, the best results are obtained by selection with the "Supervision from an Oracle" [38] paradigm (**B** selection). However, as discussed above, the ground truth is normally not available in a real-world inference phase, and then it is not completely correct to use it to select the best hypothesis. This analysis is reported here for completeness and to better illustrate the potential theoretical capabilities of the proposed method in estimating the 3D position. This result reveals how a future study on aggregation techniques could further improve the results obtained with median-based aggregation.

The superior performance of SnapPose3D is also confirmed on the hardest subset of Human3.6, here referred to as H36MA [14,52], and reported in Table 3. This evaluation can be considered an important result due to the ambiguity of the poses contained. In particular, our method outperforms all comparable methods and in particular our direct competitor [14] by a large and consistent margin of 23.5 and 15.1mm for MPJPE and P-MPJPE, respectively.

Table 4: Results on MPI-INF-3DHP dataset. Following the common literature testing protocol, ground truth 2D poses are used as input.

Method	Pavvlo [34]	Zheng [60]	Wang [51]	Li [24]	Zheng [60]	Zhang [56]	Zhao [58]	*SnapPose3D* (H=20, R)	*SnapPose3D* (H=20, M)
PCK ↑	86.0	88.6	86.9	93.8	95.4	94.4	96.8	96.1	**97.2**
AUC ↑	51.9	56.4	62.1	63.3	63.2	66.5	70.7	70.2	**71.4**
MPJPE ↓	84.0	77.1	68.1	58.0	57.7	54.9	44.7	45.5	**42.2**

Evaluation on MPI-INF-3DHP. In Table 4, we report the results in terms of PCK, AUC, and MPJPE on the test set of MPI-INF-3DHP that, differently from the previous benchmark, contains outdoor sequences. Following the testing protocol of previous works [24,56,60], we use the ground truth 2D pose detection as input. We trained the state-of-the-art competitor [58] on this dataset as well, using the publicly available implementation provided by the authors. Our approach achieves state-of-the-art results across all metrics, surpassing both non-temporal baselines and advanced temporal models, demonstrating strong generalization capabilities, especially in challenging outdoor environments.

Computational load and speed. Using batch size $B = 1$, SnapPose3D reaches an inference time ranging between 3 and 10 FPS considering up to $H = 20$ hypotheses and $K = 20$ iterations. The GPU memory occupation is 708MB on a single Nvidia RTX 3090. These results highlight how the system, although based on computationally demanding models (*e.g.*, the diffusion-based model), can still achieve near real-time performance.

Table 5: Ablation study on Human3.6M masking part of the conditioning input, reported through the MPJPE metric.

Pose Context		**Multiple Hypothesis Aggregation**			
		A (avg)	**M** (median)	**B** (Pose-level)	**B** (Joint-level)
✓		59.9	59.8	56.8	49.9
	✓	45.3	45.4	42.3	37.6
✓	✓	**43.2**	**42.8**	**36.9**	**28.6**

5.2 Ablation studies

Denoising conditioning. In Table 5, we assess the performance of the diffusion-based denoising process depending on the visual feature conditioning. We report different MPJPE metrics aggregating or selecting the hypotheses with different techniques. It is worth noting that removing the context information from the learning process degrades the performance with -27.9% on average while removing the pose features also impacts the results with -4.6% on average. This analysis of the conditioning selection demonstrates that both visual and pose

features are needed to obtain precise and reliable 3D human poses. The last column reports the performance of selecting the closest joint to the ground truth (**B**est, joint-level) across all hypotheses. Since the best joints may come from different solutions, this represents a lower bound for selection approaches.

Confidence score. The generation of multiple and different but still plausible hypotheses not only allows for the improvement of performance thanks to aggregation techniques (see Sec. 3.4), but also the definition of a confidence score. To this aim, we computed and used as the confidence score of each pose prediction the variance between the H different provided hypotheses. More precisely, we computed the average between the variances at the joint coordinate level. For example, admitting a recall of around 90% it is possible to improve the MPJPE metric by more than 2mm. To validate the chosen confidence score, we computed metrics on the most confident outputs, obtaining a reduction in global error at the cost of lower recall (i.e., reduced pose coverage), as shown in Figure 2. Notably, the confidence metric requires no ground truth, as it is computed directly on the generated outputs.

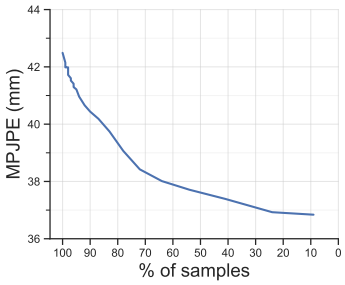

Fig. 2: Confidence analysis in terms of MPJPE

Number of generated hypotheses. To evaluate the best number of hypotheses H to generate, the MPJPE error (Protocol 1) was calculated with different aggregation/selection techniques on the Human3.6 dataset. The number of hypotheses varied between 20 and 200. The graph in Figure 3 shows the results of SnapPose3D using the aggregation methods proposed on the Human3.6 dataset. The median operator always outperforms the average one. The Oracle supervision generates the best solutions, as expected. However, median and average aggregations do not improve with more than 40 hypotheses, making $H = 20$ the best compromise between performance and computational speed. On the contrary, the selection of oracles continues to improve as the number of hypotheses increases, as can be reasonably imagined from the generation of further possibilities from which to choose the best result.

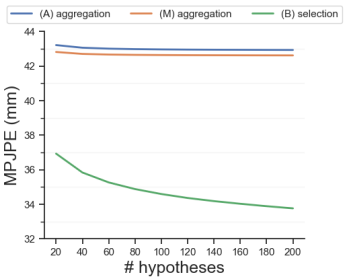

Fig. 3: MPJPE vs. number of generated hypotheses

6 Conclusion

In this work, we presented SnapPose3D, a novel diffusion-based framework for 3D human pose estimation from a single image. Unlike most previous methods that rely on temporal sequences to resolve depth ambiguity, SnapPose3D utilizes a single-frame approach, reducing the complexity of data acquisition and enabling online applications. By leveraging the generative capabilities of diffusion models, our method produces multiple plausible hypotheses for each input 2D pose and context feature. Experimental results obtained on two public benchmarks confirm that the generation and aggregation of multiple hypotheses provide significant advantages.

References

1. Arac, A., Zhao, P., Dobkin, B.H., Carmichael, S.T., Golshani, P.: Deepbehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. Frontiers in systems neuroscience **13**, 20 (2019)
2. Bishop, C.M.: Mixture density networks (1994)
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR. pp. 7103–7112 (2018)
6. Choi, J., Shim, D., Kim, H.J.: Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3773–3780 (2023)
7. D'Eusanio, A., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R.: Refinet: 3d human pose refinement with depth maps. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2320–2327. IEEE (2021)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021)
9. D'Eusanio, A., Simoni, A., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R.: Depth-based 3d human pose refinement: Evaluating the refinet framework. Pattern Recognition Letters **171**, 185–191 (2023)
10. Fan, C., Mei, Q., Li, X.: 3d pose estimation dataset and deep learning-based ergonomic risk assessment in construction. Automation in Construction **164** (2024)
11. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: CVPR. pp. 13041–13051 (2023)
12. Gower, J.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975)
13. Ho, J., Jain, A., Ab, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
14. Holmquist, K., Wandt, B.: Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: ICCV (2023)
15. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: ECCV. pp. 68–84 (2018)
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI **36**(7), 1325–1339 (2013)

17. Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In: ICCVW. pp. 805–814 (2017)
18. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV. pp. 3334–3342 (2015)
19. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2016)
22. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: CVPR. pp. 9887–9895 (2019)
23. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
24. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: CVPR. pp. 13147–13156 (2022)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
26. Liu, K., Ding, R., Zou, Z., Wang, L., Tang, W.: A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In: ECCV (2020)
27. Liu, Y., Yang, J., Gu, X., Guo, Y., Yang, G.Z.: Ego+ x: An egocentric vision system for global 3d human pose estimation and social interaction characterization. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5271–5277. IEEE (2022)
28. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV. pp. 2640–2649 (2017)
29. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV. pp. 506–516 (2017)
30. Oikarinen, T., Hannah, D., Kazerounian, S.: Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In: IJCNN. pp. 1–9 (2021)
31. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)
32. Paudel, P., Kwon, Y.J., Kim, D.H., Choi, K.H.: Industrial ergonomics risk analysis based on 3d-human pose estimation. Electronics **11**(20), 3403 (2022)
33. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR. pp. 7025–7034 (2017)
34. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR (2019)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
36. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3d human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding **152**, 1–20 (2016)

37. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: ICCV (2023)
38. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3d human pose estimation by generation and ordinal ranking. In: ICCV (2019)
39. Simo-Serra, E., Ramisa, A., Alenya, G., Torras, C., Moreno-Noguer, F.: Single image 3d human pose estimation from noisy observations. In: CVPR (2012)
40. Simoni, A., Borghi, G., Garattoni, L., Francesca, G., Vezzani, R.: D-spdh: Improving 3d robot pose estimation in sim2real scenario via depth data. IEEE Access
41. Simoni, A., Pini, S., Borghi, G., Vezzani, R.: Semi-perspective decoupled heatmaps for 3d robot pose estimation from depth maps. IEEE Robotics and Automation Letters **7**(4) (2022)
42. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. In: CVPR (2003)
43. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
44. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
45. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
46. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV. pp. 2602–2611 (2017)
47. Terreran, M., Barcellona, L., Ghidoni, S.: A general skeleton-based action and gesture recognition framework for human–robot collaboration. Robotics and Autonomous Systems **170**, 104523 (2023)
48. Vaswani, A.: Attention is all you need. NeurIPS (2017)
49. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV. pp. 601–617 (2018)
50. Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., Shao, L.: Deep 3d human pose estimation: A review. Computer Vision and Image Understanding **210** (2021)
51. Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3d pose estimation from videos. In: ECCV. pp. 764–780 (2020)
52. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows. In: ICCV. pp. 11199–11208 (2021)
53. Xu, T., Takano, W.: Graph stacked hourglass networks for 3d human pose estimation. In: CVPR. pp. 16105–16114 (2021)
54. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Adv. in neural information proces. systems **35** (2022)
55. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: CVPR. pp. 5255–5264 (2018)
56. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: CVPR (2022)
57. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: CVPR. pp. 3425–3435 (2019)
58. Zhao, Q., Zheng, C., Liu, M., Chen, C.: A single 2d pose with context is worth hundreds for 3d human pose estimation. In: NeurIPS (2023)
59. Zhao, W., Wang, W., Tian, Y.: Graformer: Graph-oriented transformer for 3d pose estimation. In: CVPR. pp. 20438–20447 (2022)
60. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: ICCV (2021)