DiffPose: Multi-hypothesis Human Pose Estimation using Diffusion Models

Karl Holmquist Bastian Wandt
Linköping University
[name.surname]@liu.se

Abstract

Traditionally, monocular 3D human pose estimation employs a machine learning model to predict the most likely 3D pose for a given input image. However, a single image can be highly ambiguous and induces multiple plausible solutions for the 2D-3D lifting step, which results in overly confident 3D pose predictors. To this end, we propose DiffPose*, a conditional diffusion model that predicts multiple hypotheses for a given input image. Compared to similar approaches, our diffusion model is straightforward and avoids intensive hyperparameter tuning, complex network structures, mode collapse, and unstable training. Moreover, we tackle the problem of over-simplification of the intermediate representation of the common two-step approaches which first estimate a distribution of 2D joint locations via joint-wise heatmaps and consecutively use their maximum argument for the 3D pose estimation step. Since such a simplification of the heatmaps removes valid information about possibly correct, though labeled unlikely, joint locations, we propose to represent the heatmaps as a set of 2D joint candidate samples. To extract information about the original distribution from these samples, we introduce our* embedding transformer *which conditions the diffusion model. Experimentally, we show that DiffPose improves upon the state of the art for multi-hypothesis pose estimation by 3-5% for simple poses and outperforms it by a large margin for highly ambiguous poses.*[1]

1. Introduction

Human pose estimation from monocular images is an open research question in computer vision with many applications, *e.g.* in human-machine interaction, autonomous driving, animation, sports, and medicine. Recent advances in deep learning-based human pose estimation show promising results on the path to highly accurate 3D reconstructions from single images. Typically, a neural network is trained to reconstruct the most likely 3D pose given an input image. However, the projection from 3D to a 2D plane, which is performed by a camera capturing a person, results in an inevitable loss of information. This lost information cannot be uniquely reconstructed, and therefore we argue that a meaningful 3D human pose estimator must be able to recover the full distribution of possible 3D poses for a given 2D pose, *e.g.* as a set of poses with different likelihoods. Moreover, downstream tasks can be built to benefit from unlikely poses; for example, consider an autonomous vehicle making decisions based on a single output versus being able to see all possible, though less likely, outcomes. Consequently, interest in this field, called multi-hypothesis human pose estimation, is increasing [16, 25, 35, 40, 23, 52, 29]. Some approaches estimate a small fixed-size set of poses [16, 25, 35, 29] which are not able to fully represent the real output distribution. Others are based on variational autoencoders [40] or normalizing flows [23, 52] and can predict an

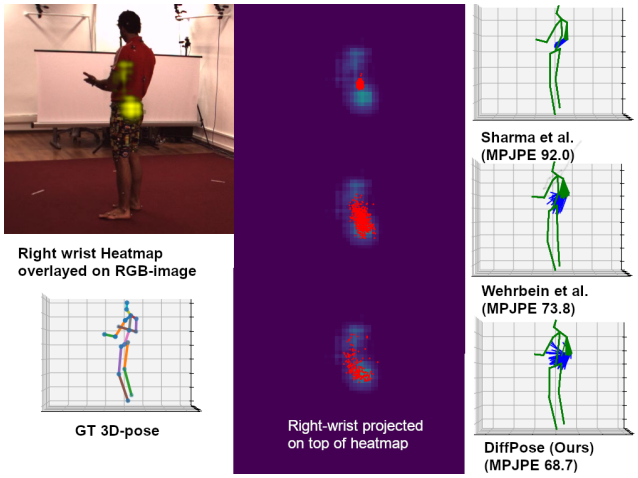

Figure 1. Comparison of our approach to Sharma *et al.* [40] and Wehrbein *et al.* [52]. While [40] produces very similar poses, even for uncertain detections, [52] achieves a higher diversity. However, they oversimplify heatmaps as a Gaussian and, thus, struggle with different uncertainty distributions. Note that the densest region of samples (red) for [40] and [52] is very similar and at a point with low certainty. By contrast, DiffPose produces 3D poses that cover the full uncertainty in the heatmap leading to a lower error.

[1] Our code and trained models will be made available at: https://github.com/bastianwandt/DiffPose/

infinite set of poses that provides a stronger approximation for the 3D pose distribution. However, they require complex architectures and lack diversity in their output, since the 2D input data is simplified, as shown in Fig. 1.

Our goal is a multi-hypothesis human pose estimator that is easy to train and produces high-quality pose hypotheses covering the full range of possible and plausible output poses. To this end, we make three major contributions: we 1) are the first to represent a 3D human pose distribution with a conditional diffusion model which, in its surprisingly simple architecture and training, achieves state-of-the-art results, 2) use the full 2D input information from heatmaps without any simplifications by our novel sampling strategy, and 3) propose a transformer architecture that handles these samples without losing information about joint uncertainties.

Neural diffusion models recently gained huge interest due to their impressive performance in image generation [37, 38, 39]. We exploit their capability to generate even subtle details that formerly were only achievable by hard-to-train GANs [49, 8] or normalizing flows [53, 23, 52, 48]. Even in its simplicity, our diffusion model creates meaningful human poses, and, unlike VAEs and GANs, it does not suffer from mode collapse, posterior collapse, vanishing gradients, and training instability [20]. Although pose representations via normalizing flows also do not show such phenomena, they require a sophisticated model of the human kinematic chain [53], a kinematic chain prior [52], and additional care during training. By contrast, our diffusion model is robust during training and creates meaningful poses without requiring further constraints.

Our second major contribution reveals a problem in current two-step approaches that first predict 2D joint positions in an image and consecutively use these predictions as input to the 3D reconstruction step. While this enables the 3D estimator to be agnostic to the input image and consequently promises generalization across image domains, it removes valid structural and depth information that can only be seen in the images. We exploit that most 2D human pose detectors employ heatmaps encoding joint occurrence probabilities as an intermediate representation. Traditionally, the maximum argument of these heatmaps is used as input to the second stage, which removes all information about the uncertainty of the detector. Few approaches extract additional information, such as confidence values [50] or Gaussian distributions fitted to the heatmap [52]. However, they still oversimplify the heatmap as shown in Fig. 1, thus missing important details. To this end, we propose to condition the diffusion model with an embedding vector computed from a set of joint positions directly sampled from the heatmaps. We build a so-called *embedding transformer* which combines joint-wise samples and their respective confidences into a single embedding vector that encodes the distribution of the joints.

2. Related Work

Monocular 3D human pose estimation is a huge field with vast and diverse approaches. Hence, this section focuses on the closest related work, namely two-stage approaches[2] and competing multi-hypothesis methods. In contrast to approaches that estimate a 3D human body shape [3, 18, 21, 22, 27, 36, 53, 55], we focus on predicting the 3D locations of a set of predefined joints.

Lifting 2D to 3D. We follow the vast body of work that estimates 3D poses from the output of a 2D pose detector [33, 6, 7, 9, 12, 14, 28, 41, 49, 50, 51, 54]. These two-stage approaches decouple the difficult problem of 3D depth estimation from the easier 2D pose localization. With the 3D lifting step being agnostic to the image data, it is easily transferable to other image domains, e.g. in-the-wild data. Moreover, in contrast to 3D training data, 2D images are significantly easier to annotate, and, therefore, a huge amount of labeled in-the-wild images are already readily available, which reduces bias towards indoor scenes that are common in 3D datasets. Early work in learning-based pose estimation is done by Akhter and Black [1] who learn a prior to restricting invalid 3D pose reconstructions. The simplest and very influential approach that commonly serves as a baseline is proposed by Martinez *et al*. [31], who employ a fully-connected residual network to lift 2D detections to 3D poses, surprisingly outperforming previous approaches by a large margin.

The above approaches predict a single most likely pose for a given input. By contrast, we predict a set of plausible 3D poses from a single 2D pose. Additionally, we leverage the full output heatmap of the 2D pose detector, which was previously simplified to its maximum, an uncertainty label [5, 50, 54], or Gaussian distributions [52]. With our novel heatmap sampling strategy, we are able to reflect the full uncertainty of the 2D predictor in our 3D pose hypotheses.

Multi-hypothesis 3D human pose estimation. Ambiguities of monocular 3D human pose estimation and sampling multiple 3D poses by heuristics are discussed in early work [24, 42, 44, 45]. Recently, few approaches have been proposed that use generative machine learning models which generate multiple diverse hypotheses to cover the ambiguous nature of 3D human pose estimation. Jahangiri and Yuille [16] uniformly sample the learned occupancy matrices [1] to generate multiple hypotheses from a predicted seed 3D pose. They use a rejection sampling approach based on a 2D reprojection error in combination with bone lengths constraints. Li and Lee [25] learn the multimodal posterior distribution using a mixture density network (MDN) [4]. They define a 3D hypothesis by the

[2] A 2D joint detection step is followed by a 3D lifting step.

conditional mean of each Gaussian kernel. Oikarinen *et al*. [35] improve [25] by utilizing the semantic graph neural network of [56]. A major limitation is the requirement of an a priori decided number of hypotheses. In contrast, Sharma *et al*. [40] condition a variational autoencoder with 2D pose detections that is capable of producing an unlimited number of hypotheses. They rank the 3D pose hypotheses by estimated joint-ordinal depth relations from the image. Kolotouros *et al*. [23] estimate parameters of the SMPL body model [30] using a conditional normalizing flow. Wehrbein *et al*. [52] also propose a normalizing flow to model the posterior distribution of 3D poses. They stabilize the training by a multitude of losses, including a pose discriminator network similar to generative adversarial networks [11]. By contrast, our diffusion-based 3D pose estimator requires only a single loss and converges stably while improving upon previous approaches, especially on a selected subset of very ambiguous poses. Moreover, we show that our approach generates more physically plausible poses. Unlike [52] we do not simplify the 2D heatmaps as a Gaussian distribution, but instead leverage the entire heatmap enabling 3D pose predictions that fully reflect the uncertainty in the 2D predictions. Li *et al*. [29] employ a transformer to learn a distribution from temporal data that is represented by 3 hypotheses which are later merged to predict a single one. In contrast, DiffPose can predict an infinite number of poses, therefore, representing the distribution more accurately and does not require temporal data.

3. Method

Our aim is to generate a set of realistic and accurate 3D human poses which approximates the full posterior distribution by utilizing a generative model. Similar to normalizing flows, which have previously been used for multi-hypothesis pose generation [52], we model the ambiguity caused by the loss of information when projecting 3D data into the image plane by conditioning a diffusion model on the 2D detections. Our model is inspired by Denoising Diffusion Probabilistic Models (DDPMs) [13] because of their recent impressive performance and stable training in image generation compared to previous generative models. We formulate the diffusion process as the iterative distortion of a vector containing 3D joint coordinates into a Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$. The denoising process is conditioned on joint-wise heatmaps that are generated by the 2D joint detector HRNet [46] using our novel embedding transformer. Fig. 2 shows the full model.

We follow the huge body of prior work by using a two-stage approach that decouples the 2D pose estimation from the 3D lifting step. With such an approach, the 3D pose estimator is agnostic to the image features and, therefore, does not overfit to a specific scenario, such as scene lighting or similar background, which is very common in motion capture datasets. However, while this behaviour is beneficial for generalization of the 3D pose estimator, it removes valuable uncertainty information when mapping from heatmaps, which most 2D pose estimators predict, to joint positions. Previous work has primarily either utilized the maximum likelihood estimate from the 2D joint detector, included confidence values for individual joints [50], or fitted a Gaussian to approximate the heatmaps [52]. However, while the heatmap for simple poses without occlusions can be well represented as a Gaussian, it can be misleading for more complex situations, *e.g*. heatmaps with multi-modal distributions that often occur for occluded joints, as shown in Fig. 1. Other common uncertain cases are left-right flips and a cluttered background. As such, we directly utilize the predicted joint position likelihoods to sample the heatmap and utilize both the samples themselves as well as their individual likelihoods to condition the reverse diffusion process.

3.1. Diffusion Model

The diffusion model consists of two parts, each defined as a Markov chain: 1) *the forward process* which iteratively adds Gaussian noise of pre-defined mean and variance to the original data, gradually distorting the data and 2) *the reverse process* which is performed by a neural network trained on a step-wise version of the degradation.

The forward process is the approximate posterior $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ modeled by a Markov chain that gradually adds Gaussian noise to the original data $\boldsymbol{x}_0$ to transform it into a Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$. It is performed by a predefined noise schedule which adds noise parameterized by β_t depending on the step t, to the original signal $\boldsymbol{x}_0$. We adopt the cosine-schedule proposed by [34], which adds a smaller amount of noise near $t = 0$ compared to a linear schedule. At each step t the noise is added incrementally to the signal according to

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}). \quad (1)$$

This formulation allows for sampling of degraded samples at any given time-step in closed form by

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\right), \quad (2)$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

The reverse process is the joint distribution $p_\theta(\boldsymbol{x}_{0:T})$ and iteratively reverts the degradation by estimating a Gaussian distribution,

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}), \Sigma_\theta(\boldsymbol{x}_t, t)). \quad (3)$$

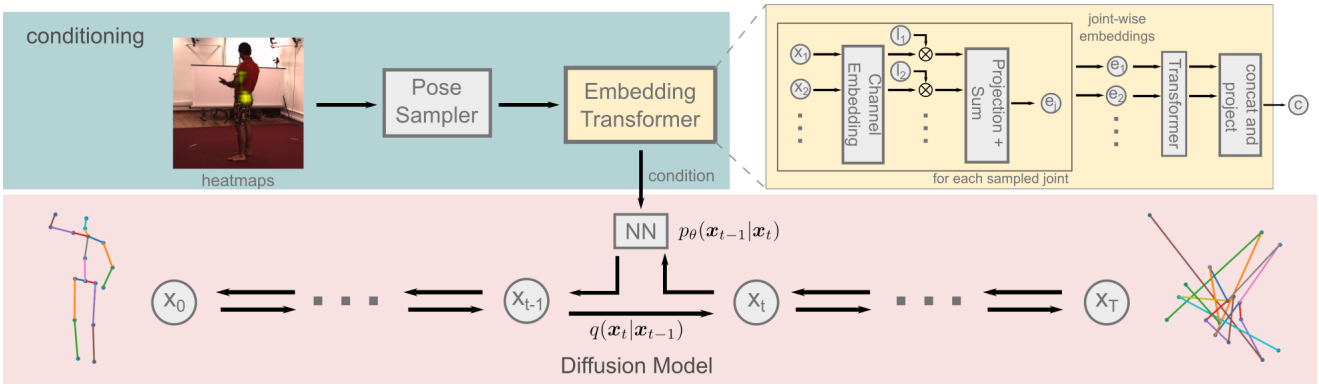

Figure 2. Overview of our proposed method. It consists of two parts: the diffusion model and the conditioning. The diffusion model alone is able to generate meaningful 3D poses. To generate multiple hypotheses for the 2D to 3D lifting process the conditioning on the 2D heatmaps in each step of the denoising process is a crucial part. Using our proposed heatmap sampling in combination with our embedding transformer that predicts an embedding for all sampled joints we achieve diverse and meaningful 3D pose predictions.

We follow DDPM by setting $\Sigma_\theta(\boldsymbol{x}_t, t) = \beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\boldsymbol{I}$ and parameterizing the predicted mean in terms of the current data $\boldsymbol{x}_t$ and the predicted noise ϵ_θ conditioned on $\boldsymbol{c}$,

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})\right). \quad (4)$$

For a derivation and more details, we refer to [13].

The noise is predicted using a neural network, parameterized by θ, which takes as input a single input vector built by concatenating the preprocessed conditioning vector $\boldsymbol{c}$ which is an embedding of the joint-wise heatmaps, the current 3D pose $\boldsymbol{x}_t$ and the current time-step t. Details about the construction of the condition vector and the exact network architecture are described in Sec. 3.3 and 3.4, respectively.

3.2. Sampling from Heatmaps

In order to represent a heatmap in a compact yet concise way, we interpret it as an independent, multinomial distribution over possible detections in a 64×64-grid (the output dimension of each heatmap from HRNet [46]) and draw n samples with replacement for each joint. For uncertain joints, *e.g.* when they are occluded, the distribution can be highly asymmetric, and previous methods struggle to approximate them as shown in Fig. 1. The sampled 2D poses are mean centered and normalized by their standard deviation. In addition to the sampled poses, we include the most likely 2D pose as one of the samples.

3.3. Conditioning the Diffusion Model

While integrating a condition into a diffusion model can be done in many ways [47, 43], there are two key aspects that need to be considered: 1) the individual joint heatmaps are independent, *i.e.* they do not contain any cross-correlation between the joints, and 2) directly averaging individual joint samples will result in a loss of the multi-modal information contained in them. We address both with our *embedding transformer* which is split into two steps as illustrated in Fig. 2. In a first step, we embed all samples for each joint non-linearly into a single vector, thus, maintaining their multi-modality. Subsequently, to account for inter-joint relationships, these embeddings are used as the input for a transformer network.

The joint-wise embedding needs to preserve the positional information and the likelihood from the heatmap of each joint sample. We use *channel embeddings* [10] to encode both into a soft histogram which maintains the positional information when averaging multiple samples. The channel embedding is created by using a kernel-basis to spread each sample over multiple of the K-bins, constituting the embedding. We utilize the truncated $\cos^2$-basis,

$$b(x) = \begin{cases} \cos^2\left(\frac{\pi x}{h}\right) & \text{for } |x| < \frac{h}{2} \\ 0 & \text{else,} \end{cases} \quad (5)$$

where the bandwidth is $h = \frac{8}{K}$, to let each basis accumulate information from all samples within a distance of 4 bins from the center location. The channel embedding is first applied to each spatial dimension independently to create a non-linear embedding which is then concatenated to a single vector per sample. Each embedding is scaled by the likelihoods of the corresponding individual joint samples, which forces the following steps to not ignore it. The scaled embedding is passed through a linear layer to introduce sample-wise *cross-spatial dependencies*. Finally, the individual joint samples are combined into a single joint embedding $\boldsymbol{e}_j$ with

$$\boldsymbol{e}_j = \sum_{n=0}^{N} \text{MLP}\left(l^n \begin{bmatrix} b\left(x_x^n + \frac{2s}{K}\right) \\ b\left(x_y^n + \frac{2s}{K}\right) \end{bmatrix}_{s=-\frac{K}{2}}^{\frac{K}{2}}\right), \quad (6)$$

where l^n is the likelihood of the sampled joint position.

Inter-joint dependencies are introduced by adding learned positional encodings to the embeddings in order to distinguish the joints and passing them to a transformer network. The outputs of the transformer are joint-wise embeddings that can now also contain information about other joints. To create the final combined conditioning vector $\boldsymbol{c}$, the embeddings are concatenated and passed through a linear projection layer.

3.4. Implementation Details

Optimization. Both the denoiser and the conditioning are optimized jointly by minimizing the simplified loss objective from Ho *et al*. [13]

$$\mathcal{L} := \mathbb{E}_{t,\boldsymbol{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, \boldsymbol{c})\|^2\right] \quad (7)$$

sampled at uniform time steps, t, where $\epsilon \sim \mathcal{N}(0, I)$. In contrast to previous work in multi-hypothesis 3D human pose estimation, we only require a single loss term and one neural network, which makes training simple and stable for a wide range of hyperparameters.

The denoiser is a linear layer followed by two residual blocks, each containing two fully connected layers of dimension 1024 with a LeakyReLU as activation function, similar to [31]. For inference efficiency, our proposed method only samples the heatmap and calculates the condition vector once per forward-pass instead of generating new samples at each time-step t.

2D detector. We use the state-of-the-art and publicly available model HRNet [46], pretrained on MPII [2] and fine-tuned on Human3.6M [52]. Although any model that produces a heatmap of each individual joint would be possible to use, we chose this one specifically for comparability with previous methods [52]. For completeness we also include the results based on the non-finetuned HRNet model in the supplementary.

Data preprocessing. The raw 64×64-pixel heatmaps are generated from cropped square regions as in [52]. The sampled 2D joint positions are mean-centered and normalized by the standard deviation. The 3D poses are processed in decimeters and mean centered individually.

Training. The network is trained for 700k iterations using Adam [19], a learning rate of 1×10^{-4}, and a batch size of 64. We set $K = 64$ for the channel embedding and project the final condition into a $2 \times 64 \times J = 2048$-dimensional vector before concatenating it with the time step t and $\boldsymbol{x}_{t-1}$. During training, we randomly drop individual joints by setting the joint embedding, $\boldsymbol{e}_j = 0$, with a fixed probability of 0.01. We noticed that this further improves the symmetry of generated poses and decreases the PA-MPJPE on H36MA (cf. Tab. 4). Training on a single NVIDIA A40 takes approximately 7 hours.

4. Experiments

Following previous work, we evaluate our method on the well-known benchmark datasets Human3.6M [15] and MPI-INF-3DHP [32] using their established training and test splits. For the Human3.6M dataset, we follow standard protocols and evaluate on every 64th frame of the test set.

Since our main focus is highly ambiguous poses, we evaluate on the H36MA subset of Human3.6M as defined by Wehrbein *et al*. [52]. It contains data samples where at least one Gaussian that is fitted to the heatmaps has a standard deviation larger than 5 px. This subset contains 6.4% of all samples present in the Human3.6M test set. These samples are extremely challenging since the joint detector gives inaccurate or wrong results. The results on this dataset can be seen as the main target of our approach.

In addition to Human3.6M and MPI-INF-3DHP, we use the Leeds Sports Pose extended (LSPe) dataset [17] for qualitative evaluation.

Metrics. For Human3.6M we follow the standard protocols. Protocol I calculates the mean Euclidean distance between the root-aligned reconstructed poses and ground truth joint coordinates which is commonly known as *mean per joint position error* (MPJPE). Protocol II first employs a Procrustes alignment between the poses before calculating the MPJPE, also known as PA-MPJPE. For 3DHP, we additionally report the *Percentage of Correct Keypoints* (PCK). It is the percentage of predicted joints that are within a distance of $150mm$ or less from their corresponding ground truth joint. Following Wandt et al. [50] we additionally evaluate the Correct Poses Score (CPS) which, unlike the PCK, classifies a pose as correct if all joints of the pose are correctly estimated for a given threshold, therefore, yielding a stronger metric than the PCK. To be independent of a threshold value, the CPS calculates the area under the curve in a range from $1mm$ to $300mm$. Compared to most prior work that reports the performance of a single model, we report the mean and variance over five different random seeds for each metric.

4.1. Quantitative Evaluation

We report metrics for the best 3D pose hypothesis generated by our network, which is in line with previous work. This evaluation reflects how well the learned 3D poses cover the actual ground truth distribution, which is particularly

interesting for ambiguous examples. Therefore, instead of validating whether the predictions are equal to a specific solution, we assess whether that specific solution is contained in the set of predictions. Note that we do not aim to predict a single best pose but instead predict a set of poses that approximates the posterior distribution, which enables downstream tasks to consider the uncertainty of the predictions.

Evaluation on Human3.6M. Following [40] and subsequent work, we produce $M = 200$ hypotheses for each image. The samples drawn from the heatmaps corresponding to one image remain constant for all 200 hypotheses. Table 1 compares our approach with others and shows that we improve upon the state of the art by 3.2% and 4.9% for protocols 1 and 2, respectively. Note that we match Li *et al*. [29] in MPJPE and outperform them by 10.5% in PA-MPJPE although they use temporal data.

However, our main target is highly ambiguous cases. Therefore, our core result is the evaluation on H36MA, the hard subset of Human3.6M, shown in Table 2. On average, we significantly outperform the state of the art by $8.0mm$ (11.1%) and $7.5mm$ (13.8%) in MPJPE and PA-MPJPE, respectively. Furthermore, we improve the PCK by 1.6% and the CPS by 24.5. Figure 1 shows an example of the increased diversity that results in predictions closer to the ground truth, leading to these large improvements.

Generalization to other datasets. We evaluate on the MPI-INF-3DHP dataset to show the generalization abilities of our model. The 2D detector and the diffusion model remain the same as for the Human3.6M dataset and are not trained or refined for the experiments in this section. Table 3 shows that, on average, we perform on par with the closest competitor Li *et al*. [26] as shown in the last column. For challenging outdoor scenes (column *Outdoor*) our method improves by 5.4% and 1.3% on [26] and [52], respectively. Similarly to the results on H36MA, this highlights that DiffPose is well suited for more complicated scenes. Although we follow previous work and only evaluate 200 3D pose hypotheses, an increased number of hypotheses significantly improves performance, as also shown in Fig. 4. The performance reaches a PCK of 87.6 at 2000 hypotheses, which is a large improvement over our result with 200 hypotheses and thus also over state of the art. Results for other metrics for the 3DHP dataset are reported in the supplemental document.

4.2. Qualitative Evaluation

Fig. 5 shows visual results of our method for 3 different datasets, Human3.6M, MPI-INF-3DHP, and LSPe. For better visibility we only show 10 pose hypotheses with the middle one in a stronger color. Note that the variance for visible joints in common poses is low, whereas rare poses with occluded joints show a high variance in the reconstructions. Even for MPI-INF-3DHP and LSPe we achieve plausible reconstructions although these datasets were not used for training. In cases where the reconstructed poses do not completely match the ground truth, they still have plausible joint angle limits and bone lengths, as also discussed in the ablation studies in Sec. 4.3. Occasional failure cases occur when joints are misdetected by the 2D joint detector (top, right column), or poses are too far outside of the distribution of the poses in the training dataset (bottom right).

4.3. Ablation Study

We perform several ablation studies to evaluate our method in different settings and validate our contributions.

Why diffusion models? Although diffusion models have shown amazing results for highly detailed image generation, little is known about their capabilities to model human skeletons. To verify that diffusion models are also capable to represent features at a higher abstraction level for humans, we calculate a symmetry error as the mean bone lengths difference between the left and right sides of the human body. Table 2 shows the results in the column *Sym*. Although [52] uses a kinematic prior that encourages symmetry, we achieve a significantly lower error (12.5mm or 46%) which means our generated poses are more plausible. We also outperform our closest competitor, according to the symmetry metric, [40] by $9mm$ or 38%. This is also reflected in the significantly lower PA-MPJPE shown in Tab. 2. Additionally, the results of the *Diffusion baselines* in Tab. 4 highlights that even a relatively simple diffusion model outperforms the previous state of the art on H36MA.

Number of diffusion steps. The forward and backward pass in a diffusion model is defined as a Markov process. To ensure that the forward process results in a Gaussian distribution, infinitely many steps are required. Commonly, this is approximated with a large finite number of steps. Fig. 4 shows the performance of our model for different numbers of total diffusion steps. Since fewer steps allows for faster sampling and 25 appears to be close to the optimal value we perform all our experiments with that number of steps. A larger number of steps has a slight impact on the results but it remains constant across a wide range of values while still being significantly below our closest competitor.

Embedding transformer. Tab. 4 shows the performance of our model using different ways to compute the condition vector. The possibly simplest condition is directly using the maximum argument of the heatmaps as the condition shown in row *Diffusion baseline - sample-free*. To represent the heatmap via samples, one could also use those directly

Table 1. Results in millimeters for the H36M dataset for protocol 1 (MPJPE) and protocol 2 (PA-MPJPE). The row marked with dagger † uses temporal information and is included for conciseness but not marked in bold even if it shows the best performance for some activities. We report the average metric of five random seeds for our method and the variance.

Protocol 1 (MPJPE)	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Martinez *et al.* [31] ($M = 1$)	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
Li *et al.* [26] ($M = 10$)	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
Li *et al.* [25] ($M = 5$)	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
Oikarinen *et al.* [35] ($M = 200$)	40.0	43.2	41.0	43.4	50.0	53.6	40.1	41.4	52.6	67.3	48.1	44.2	**44.9**	39.5	40.2	46.2
Sharma *et al.* [40] ($M = 10$)	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
†MHFormer Li *et al.* [29] ($M = 3$)	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
Wehrbein *et al.* [52] ($M = 200$)	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
DiffPose (Ours $M = 200$)	**37.7**	**41.5**	**38.7**	**41.6**	**45.9**	51.9	**38.8**	**37.8**	**48.4**	**54.0**	**43.3**	**44.1**	46.6	**37.8**	**34.9**	**42.9**$^{\pm 0.3}$
Protocol 2 (PA-MPJPE)	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Martinez *et al.* [31] ($M = 1$)	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
Li *et al.* [26] ($M = 10$)	38.5	41.7	39.6	45.2	45.8	46.5	37.8	42.7	52.4	62.9	45.3	40.9	45.3	38.6	38.4	44.3
Li *et al.* [25] ($M = 5$)	35.5	39.8	41.3	42.3	46.0	48.9	36.9	37.3	51.0	60.6	44.9	40.2	44.1	33.1	36.9	42.6
Oikarinen *et al.* [35] ($M = 200$)	30.8	34.7	33.6	34.2	39.6	42.2	31.0	31.9	42.9	53.5	38.1	34.1	38.0	29.6	31.1	36.3
*Sharma *et al.* [40]($M = 200$)	30.6	34.6	35.7	36.4	41.2	43.6	31.8	31.5	46.2	49.7	39.7	35.8	39.6	29.7	32.8	37.3
†MHFormer Li *et al.* [29] ($M = 3$)	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
Wehrbein *et al.* [52] ($M = 200$)	27.9	31.4	29.7	30.2	34.9	37.1	**27.3**	28.2	39.0	46.1	34.2	32.3	33.6	**26.1**	27.5	32.4
DiffPose (Ours $M = 200$)	**26.9**	**30.1**	**29.5**	**29.4**	**32.1**	**35.6**	27.7	**27.1**	**36.4**	**41.3**	**32.1**	**29.9**	**32.2**	26.7	**24.7**	**30.8**$^{\pm 0.05}$

Table 2. Results for the hard subset H36MA as defined by Wehrbein *et al.* [52]. We outperform all comparable methods by a large margin. Additionally, the symmetry error shows that in average DiffPose produces more plausible poses.

Method	MPJPE ↓	PA-MPJPE ↓	PCK ↑	CPS ↑	Sym ↓
Li *et al.* [25]	81.1	66.0	85.7	119.9	-
Sharma *et al.* [40]	78.3	61.1	88.5	136.4	23.9
Wehrbein *et al.* [52]	71.0	54.2	93.4	171.0	27.4
DiffPose (Ours)	**63.1**$^{\pm 0.4}$	**46.7**$^{\pm 0.1}$	**94.9**$^{\pm 0.01}$	**195.5**$^{\pm 3.5}$	**14.9**$^{\pm 0.02}$

Table 3. Quantitative results on MPI-INF-3DHP. We outperform all comparable methods which indicates a good generalizability of DiffPose to other sequences without requiring additional training.

Method	Studio GS ↑	Studio no GS ↑	Outdoor ↑	All PCK ↑
Li *et al.* [26]	86.9	**86.6**	79.3	**85.0**
Li *et al.* [25]	70.1	68.2	66.6	67.9
Wehrbein *et al.* [52]	86.6	82.8	82.5	84.3
DiffPose (Ours)	**87.4**$^{\pm 0.4}$	82.7$^{\pm 0.2}$	**83.6**$^{\pm 0.3}$	84.7$^{\pm 0.1}$

Table 4. Ablation study for different configurations of DiffPose on H36MA. Each of our contributions leads to clear improvements.

Configuration	MPJPE ↓	PA-MPJPE ↓	PCK ↑	CPS ↑	Sym ↓
Diffusion baseline - sample-free	67.7	50.1	93.5	178.4	25.7
Diffusion baseline - sample-based	66.4	50.0	93.5	180.4	26.9
Diffusion baseline - full heatmap	173.8	283.3	37.3	24.1	13.4
DiffPose w/o sampling	67.7	50.1	93.5	178.9	21.8
w/o maximum-likelihood sample	62.5	46.2	95.2	196.9	14.6
w/o cross-spatial dependence	63.6	46.8	94.9	194.9	14.5
w/o likelihood scaling	62.4	46.8	95.0	194.3	14.3
w/o transformer	66.0	49.2	94.2	191.6	16.0
w/o dropout	65.0	47.9	94.2	188.7	17.4
DiffPose (Ours)	63.1	46.7	94.9	195.5	14.9

as the condition by ordering them according to likelihood and concatenating them (row *Diffusion baseline - sample-based*). An alternative to our embedding is to embed the entire heatmap using a ResNet18-network as seen in row *Diffusion baseline - full heatmap*. However, these simple conditions perform significantly worse than our full model.

The remaining rows show the performance when different parts of the embedding transformer are removed. Removing 1) the MLP in Eq. (6), which combines the channel embedded x- and y-position of the samples (*w/o cross-spatial dependency*), 2) the likelihood scaling by setting $l^n = 1$ in Eq. (6) (*w/o likelihood scaling*), 3) the random dropout of joints during training (*w/o dropout*), and 4) replacing the transformer by directly concatenating the joint embeddings in Fig. 2 (*w/o transformer*) all lead to a slightly worse performance compared to our full model.

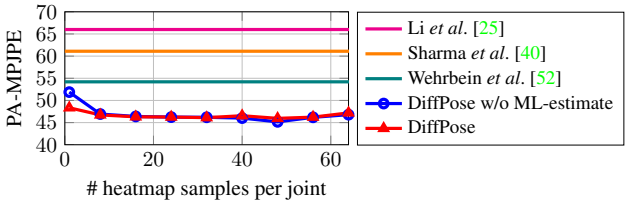

Figure 3. Evaluation results on the H36MA dataset for an increasing number of joint samples drawn from the heatmaps.

The number of joint samples drawn from the heatmap plays a crucial role for the representative power of the embedding created by the embedding transformer. Fig. 3 shows the performance for different numbers of samples per joint. Although a single sample is not enough, as also shown in Tab. 4 (row *w/o sampling*), the performance increases with more samples. We choose 32 samples in our main experiments as a good trade-off between performance and complexity. Note that the performance remains stable over a wide range of values indicating the robustness of our method against different choices of hyperparameters. Fig. 3 shows that removing the maximum likelihood (*w/o maximum likelihood sample* in Tab. 4), the performance deteriorates for a small number of joint-samples. However, the difference decreases as the number of samples increases. This underlines that our embedding transformer indeed learns to represent the full heatmap. In any configuration, we outperform other approaches with a large margin. Note that the number of samples does not influence the number of 3D pose hypotheses used for evaluation.

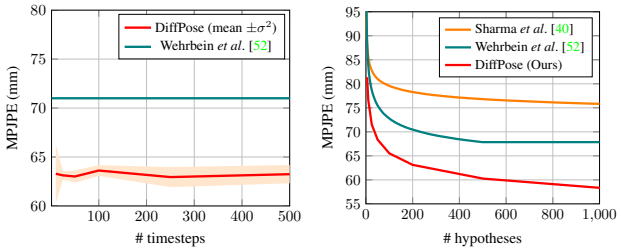

Figure 4. Evaluation results on the subset H36MA. **Left**: increasing number of timesteps in the denoising process. **Right**: increasing number of generated 3D pose hypotheses.

Number of hypotheses. Fig. 4 shows the performance on H36MA for an increasing number of 3D pose hypotheses compared to others. As expected with more hypotheses, the errors decrease. Notably, ours continues to improve for more than 1000 hypotheses which is not the case for [52] that appears to saturate around 500 hypotheses. For 2000 hypotheses, we reach an MPJPE of $56.5mm$ and a PA-MPJPE of $41.2mm$, which is significantly below the results reported in Tab. 2.

5. Limitations

In general, all two-step approaches remove image information in favor of being agnostic to the image domain, *e.g.* indoor/outdoor, lighting, and image size. Although we effectively extract more information from the heatmaps, as any other two-step approach, image information is still ignored, which could possibly be used to further refine results. However, directly incorporating it into current pose estimation methods mostly leads to degraded performance. Therefore, we still strongly advocate two-stage approaches and encourage the extraction of other valuable features from the images for further research. In the extreme case, the predicted heatmaps are entirely wrong. Fig. 5 shows that our model is only partially able to correct for these mistakes, since it tries to generate plausible poses in terms of joint angle limits and bone lengths by the strong representational power of the diffusion model. However, we are not aware of any other two-step approach that can handle these errors. We argue that these errors already arise from uncommon or difficult scenarios in the image domain, which are only reflected in the heatmaps and cannot be easily solved by including image information in the 3D lifting step.

6. Conclusion

We presented DiffPose, a conditional diffusion model that estimates multiple hypotheses for 3D human pose estimation from a single image. Our diffusion model learns plausible human poses, *e.g.* in terms of symmetry, that are valid solutions for a given input image, not only outperforming previous methods by a large margin for highly ambiguous poses, but also being simpler and more robust to train using only a single loss term. Additionally, we propose a novel sampling method from 2D joint heatmaps in combination with an embedding transformer to represent the uncertainties in the heatmaps. We show that the embeddings predicted by the transformer are superior to simpler embeddings used in prior work. We hope that our novel embedding method enables future research to use the full information in 2D joint heatmaps.

Our accurate 3D pose estimates have a wide range of applications in downstream tasks, such as 3D pose tracking, multi-view pose estimation, and likelihood estimation for pose forecasting.

Acknowledgement

This work was partially supported by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation and the Swedish Research Council grant 2018-04673. The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant 2022-06725.

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[3] Benjamin Biggs, Sébastien Ehrhadt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[4] Christopher M. Bishop. Mixture density networks. Technical report, Aston University, 1994. 2

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[6] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[7] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[8] Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial

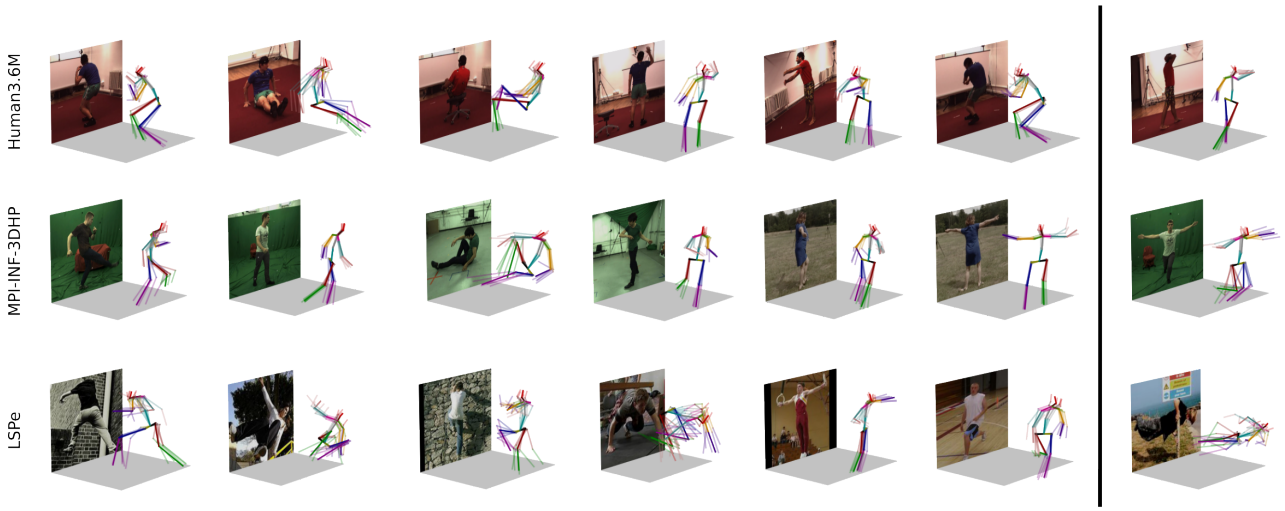

Figure 5. Qualitative results for different datasets. We achieve plausible 3D poses for a large variety of poses. The right most column shows occasional failure cases with misdetected joints (top) and poses far outside the distribution of poses in the training dataset (middle and bottom). For better visibility we only show a subset of the reconstructed poses.

parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10997–11005, 2022. 2

[9] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[10] Michael Felsberg, Kristoffer Öfjäll, and Reiner Lenz. Unbiased decoding of biologically motivated visual feature descriptors. *Frontiers in Robotics and AI*, 2:20, 2015. 4

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[12] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 4, 5

[14] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d pose estimation. *European Conference on Computer Vision (ECCV)*, 2018. 2

[15] 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 (PAMI)*, 36(7), 2014. 5

[16] Ehsan Jahangiri and Alan L. Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. *International Conference on Computer Vision Workshops (ICCVW)*, 2017. 1, 2

[17] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR) 2011*, 2011. 5

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[20] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[23] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021. 1, 2, 3

[24] Mun Wai Lee and Isaac Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004. 2

[25] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network.

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 7

[26] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3d human pose hypotheses. *British Machine Vision Conference (BMVC)*, 2020. 6, 7

[27] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 2

[28] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[29] 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. 1, 3, 6, 7

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[31] 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 (ICCV)*, 2017. 2, 5, 7

[32] 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 *International Conference on 3D Vision (3DV)*, 2017. 5

[33] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[34] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[35] Tuomas P. Oikarinen, Daniel C. Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. *arXiv preprint arXiv:2010.13668*, 2020. 1, 3, 7

[36] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[38] 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

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[40] 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 International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 6, 7, 8

[41] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), 2020. 2

[42] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[43] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021. 4

[44] Cristian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3d body tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 2

[45] Cristian Sminchisescu and Bill Triggs. Kinematic jump processes for monocular 3d human tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 2

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4, 5

[47] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021. 4

[48] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2022. 2

[49] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[50] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[51] Jue Wang, Shaoli Huang, Xinchao Wang, and Dacheng Tao. Not all parts are created equal: 3d pose estimation by modelling bi-directional dependencies of body parts. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[52] 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, 8

[53] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & Ghuml: Generative 3d human shape and articulated pose models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[54] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[55] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *European Conference on Computer Vision (ECCV)*, 2020. 2

[56] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3