Back to Optimization: Diffusion-based Zero-Shot 3D Human Pose Estimation

Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, Jenq-Neng Hwang
University of Washington, University of Copenhagen
zyjiang, zhouz47, wchai, cycyang, hwang}@uw.edu, lilei@di.ku.dk

Abstract

Learning-based methods have dominated the 3D human pose estimation (HPE) tasks with significantly better performance in most benchmarks than traditional optimization-based methods. Nonetheless, 3D HPE in the wild is still the biggest challenge for learning-based models, whether with 2D-3D lifting, image-to-3D, or diffusion-based methods, since the trained networks implicitly learn camera intrinsic parameters and domain-based 3D human pose distributions and estimate poses by statistical average. On the other hand, the optimization-based methods estimate results case-by-case, which can predict more diverse and sophisticated human poses in the wild. By combining the advantages of optimization-based and learning-based methods, we propose the Zero-shot Diffusion-based Optimization (ZeDO) pipeline for 3D HPE to solve the problem of cross-domain and in-the-wild 3D HPE. Our multi-hypothesis ZeDO achieves state-of-the-art (SOTA) performance on Human3.6M, with minMPJPE 51.4mm, without training with any 2D-3D or image-3D pairs. Moreover, our single-hypothesis ZeDO achieves SOTA performance on 3DPW dataset with PA-MPJPE 40.3mm on cross-dataset evaluation, which even outperforms learning-based methods trained on 3DPW. Our code is available here: https://github.com/ipl-uw/ZeDO-Release.

1. Introduction

As people become increasingly interested in Virtual Reality (VR), Augmented Reality (AR), Human-Computer Interaction and Sports Analysis, 3D Human Pose Estimation (HPE) becomes a crucial component for these applications. Compared with multi-view 3D HPE, monocular-based methods are easier to set up and have lower costs, which are more suitable for VR, AR, and mobile devices. Weng et al. [60], and Peng et al. [40] utilize 3D human poses with Neural Radiance Fields (NeRF) for 3D Human Reconstruction. Meanwhile, Bridgeman et al. [2] propose a 3D HPE and tracking pipeline for soccer analysis, and Jiang et al. [19] take advantage of 2D and 3D HPE to track the motion of golf players.

With the availability of more benchmark datasets, deep learning-based 3D HPE methods have been shown to outperform all traditional methods and dominate the areas. Combining 2D HPE with SMPL [30] model, Bogo et al. propose SMPLify [1] as an optimization-based 3D HPE pipeline. 2D-3D lifting [7,39,61,64,65] and diffusion-based 3D HPE [8,10] networks leverage 3D human poses from the single-frame or multi-frame 2D poses. On the other hand, Image-to-3D networks [22,24,29,52] estimate 3D human poses directly from images without intermediate 2D human poses. However, as mentioned in [9,11], learning-based 3D HPE methods suffer from performance degradation with cross-domain or in-the-wild scenarios. During training, these methods implicitly learn camera intrinsic parameters, domain-based 3D human pose distributions, or image features in a certain domain. Although optimization-based methods can mitigate the impact of domain gaps by estimating 3D poses case-by-case, their performances are not comparable to learning-based methods at this moment.

To address this problem, Zhan et al. [62] decouple the camera intrinsic parameters from 2D-3D lifting networks...
learning by converting 2D keypoints to 2D rays. Gong et al. [11] and Gholami et al. [9] generate various 3D poses to bridge domain gaps. Furthermore, Chai et al. [5] propose a data augmentation pipeline to minimize the 3D human pose spatial distribution gap. However, the above methods still cannot outperform the learning-based 3D HPE. To decouple camera intrinsic and bridge 3D pose domain gaps simultaneously, we propose the Zero-shot Diffusion-based Optimization (ZeDO) pipeline for 3D human pose estimation, which combines a simple yet effective optimization pipeline with a pre-trained diffusion-based 3D human pose generation model.

Different from traditional optimization-based methods [1, 13, 37, 54], which include various kinematic constraints, the diffusion model denoises the output from the optimization pipeline iteratively to ensure optimized poses following human body constraints. Meanwhile, poses are optimized by minimizing 2D keypoint re-projection errors with a simple yet effective optimization pipeline. Thus, ZeDO is able to estimate 3D human poses without training on any 2D-3D or image-3D pairs. Our contributions are as follows:

- The proposed ZeDO pipeline is a Zero-Shot 3D Human Pose Estimation pipeline, which leverages a pre-trained diffusion-based 3D human pose generation model to optimize target 3D poses in the loop during the inference time.

- Compared with other generation and denoising tasks, we take the diffusion model as an optimization tool by combining a simple 3D HPE optimization pipeline with a pre-trained diffusion-based 3D human pose generation model.

- ZeDO achieves state-of-the-art zero-shot 3D HPE performance on Human3.6M, 3DHP, and 3DPW datasets, even on cross-dataset evaluation.

In Sec 2, we will discuss related works in 3D HPE, including optimization-based and learning-based methods. In Sec 3, details of our backbone architecture and optimization pipeline are addressed. Experimental results are presented in Sec 4, and the ablation studies will be discussed in Sec 5. At last, there are conclusions in Sec 6.

2. Related Works

2.1. Optimization-based Human Pose Estimation

Optimization-based methods, which estimate 3D poses frame-by-frame and case-by-case, are not handicapped by domain gaps or varying camera intrinsic parameters, but their performance so far is much worse than learning-based methods. Utilizing the SMPL [30] model, SMPLify [1] is capable of optimizing the 3D human poses by minimizing the 2D keypoint re-projection error and satisfying lots of kinematic constraints. Furthermore, Müller et al. [37] propose SMPLify-XMC as an improved version of SMPLify with more constraints about the human body and more inputs including height and age. Zheng et al. [54] propose an optimization-based hierarchical 3D human pose estimation pipeline that can estimate both 3D human pose and locations at the same time. Recently, Song et al. [47] and Choutas et al. [6] focus on train an optimizer to fit the SMPL model to estimated 2D human poses.

2.2. 2D-3D Lifting Networks

2D-3D lifting networks employ either a single frame or a sequence of normalized 2D keypoints as input to generate corresponding 3D keypoints [33, 39, 61, 64]. Pavllo et al. [39] leverage dilated temporal convolutions with semi-supervised ways to improve 3D pose estimation in videos. Yang et al. [61] enhance 3D human pose estimation by leveraging in-the-wild 2D annotations and a novel refinement network module in a weakly-supervised framework. Li et al. [26] propose a scalable data augmentation technique that synthesizes unseen 3D human skeletons for training 2D-to-3D networks, effectively reducing dataset bias and improving model generalization to rare poses. These methods, despite the need of two-stage processing to obtain the 2D keypoints in advance, have demonstrated superior performance on several benchmark datasets and are highly efficient, especially when adapted for temporal considerations.

2.3. Image-to-3D Networks

End-to-end 3D HPE methods directly transform image data into 3D pose representations, such as those put forth by Guler et al. [14], Tung et al. [56], Tan et al. [53], and various other research teams including those behind SPIN [24], ROMP [51], BEV [52], and CLIFF [29], who successfully utilize scale and variable height information, effectively resolving issues of depth/height ambiguity. For instance, the methodology introduced by Sun et al. [51] is a one-stage process that allows for real-time, monocular 3D mesh recovery of multiple individuals. Further contributing to this field, Sun et al. [52] develop a single-shot method capable of simultaneously regressing the pose, shape, and relative depth of multiple people within a single image, utilizing the Bird’s-Eye-View representation for depth reasoning while accommodating variable heights. Within the area of 3D pose estimation from single images, these one-stage techniques consistently demonstrate robust performance, despite their comparatively streamlined architectural designs.

2.4. Diffusion Models in Human Pose Estimation

Nowadays, Diffusion Probabilistic Model [46] and its descendants [15, 44, 48] have shown their outstanding capabilities of prior distribution learning in multiple areas,
such as image generation [42] and editing [3, 4], 3D-reconstruction [41], image inpainting [31], and human motion generation [55]. Intuitively, diffusion models are able to tackle depth/scale ambiguity and one-to-many mappings in 3D HPE tasks. Ci et al. [8] introduce a novel score-based generative framework to model plausible 3D human poses with a hierarchical condition masking strategy. Meanwhile, Gong et al. [10] propose a diffusion-based architecture including the initialization of 3D pose distribution, a GMM-based forward diffusion process, and a conditional reverse diffusion process.

Although learning-based methods achieve significant success in 3D HPE tasks, they are highly constrained by the training datasets. In other words, their performance would abruptly drop when tested on other datasets due to domain gaps. To resolve this issue, we integrate the generality of optimization methods into a robust pre-trained diffusion backbone and propose a Zero-shot Diffusion-based Optimization pipeline (ZeDO) for 3D HPE.

3. Method

As shown in Fig 2, ZeDO includes an initial pose optimizer for rotating initial poses and an optimizer in the loop for iteratively optimizing 3D poses.

Firstly, a randomly selected initial 3D pose (a hypothesis), \( P_{\text{init}} \in \mathbb{R}^{T \times 3} \), is rotated to an optimal pose, \( P_0 \), by minimizing the re-projection error with detected or ground truth 2D keypoints \( P_{2d} \in \mathbb{R}^{J \times 2} \). Here, \( J \) stands for the number of keypoints. Then, in the \( i \)th optimization step, \( P_i \) is optimized by the optimizer in the loop, and the pre-trained diffusion model is used to denoise it to \( P_{i+1} \) as input of the next iteration. After \( n \) iterations, \( P_n \) will be the estimated 3D human pose.

Different from other diffusion-based pose estimation methods [8, 10], our diffusion model \( \theta_g \) is a pose generation model, which is only trained with 3D human poses, and during inference, our diffusion model only takes the optimized pose \( P_i \) and timestamp \( t(i) \) as input without any additional pose condition information including 2D poses.

3.1. Pre-trained 3D Human Pose Generation Model

We apply the Score Matching [50] on the pre-trained backbone for our 3D human pose generation diffusion model, which rectifies the noisy poses generated after projection to get reasonable 3D poses. During pre-training, the model takes relative-to-pelvis 3D poses \( x \in \mathbb{R}^{J \times 3} \) as inputs and tries to recover them from recurrent Gaussian noise. In this case, we expect that the diffusion model learns the distribution of real 3D poses and reconstructs \( \tilde{x} \in \mathbb{R}^{J \times 3} \) to minimize the difference from the inputs. The perturbation strategy used in our Score Matching diffusion model expresses \( p(x(t)|x(0)) \) in the closed form as:

\[
N \left( x(t); x(0)e^{-\frac{1}{2} \int_0^t \beta(s) ds}, \left[ 1 - e^{-\int_0^t \beta(s) ds} \right]^2 I \right). \tag{1}
\]

Besides, built upon the learning strategy of the noise conditional score network (NCSN) [49], we formulate our loss

![Figure 2. The pipeline of ZeDO, which takes an initial 3D pose, called a hypothesis, as input and estimates the pose by minimizing re-projection error with the target 2D pose. After 1000 iterations, ZeDO is able to generate the optimized 3D human poses.](image_url)
Before

\[ P_0 \]

After

\[ P_0 \quad P_{100} \quad P_{1000} \]

Figure 3. The rotation optimization in the initial pose optimizer finds the optimal initial pose and prevents the collapse of the estimated pose.

function as follows by choosing \( \lambda(t) = \sigma(t)^2 \):

\[
L = E_{U(t;0,1)} \left[ \lambda(t) \| \sigma(x(t), t) + \frac{x(t) - \mu}{\sigma^2} \|_2^2 \right] \quad (2)
\]

\[
= E_{U(t;0,1)} \left[ \| \sigma(x(t), t) + z \|_2^2 \right], \quad (3)
\]

in which \( z \) stands for random noise vector \( z \sim N(0,1) \) and \( \sigma \) is the pre-trained score matching network. \( \sigma \) represents the variance mentioned in Eq.(1) as \([1 - e^{-\int_0^t \beta(s)ds}]\) and the timestamp or denoising time variable \( t \) is uniformly sampled 1000 times from \((0,1)\). The 3D pose generation model is never trained with any 2D-3D or image-3D pairs. We report the results based on the Score Matching based model, similar to GFPose [8]. More results with different backbones, including DDIM [48] and DDPM [16], are shown in Table 5.

3.2. Initial Pose Optimizer

Similar to other optimization-based 3D HPE methods [1, 13, 54], our optimization pipeline starts from an initial pose, \( P_{init} \). As depicted in Fig 3, the optimized pose may not suffice if the initial pose’s orientation is significantly different from that of the target pose or even perpendicular to that of the target pose. Therefore, the initial pose optimizer is designed to find the optimal rotation matrix, \( R_0 \in SO(3) \), and translation, \( T_0 \), of \( P_{init} \) by minimizing the re-projection error with 2D keypoints \( p_{2d} \).

\[
\arg \min_{R_0, T_0} \left\| K(R_0 P_{init} + T_0) - p_{2d} \right\|_2 \quad (1)
\]

\[
s.t. \quad T_{min} \leq T_0 \leq T_{max}, \quad (2)
\]

where \( K \) is the camera intrinsic matrix. After the rotation, the \( P_0 = R_0 P_{init} \) is the optimal pose sent to the iterative optimization pipeline. As shown in Fig 3, the rotation optimization aligns the initial poses with target 2D and 3D poses.

Algorithm 1 ZeDO pipeline

Require: Initial 3D pose \( P_{init} \), Target 2D pose \( p_{2d} \), 2D pose confidence scores \( C_{2d} \), Camera intrinsic \( K \), Diffusion timestamp \( t \), Pre-trained diffusion model \( \theta_g(P,t) \)

\( R_0, T_0 \leftarrow \arg \min_{R_0, T_0} \left\| K(R_0 P_{init} + T_0) - p_{2d} \right\|_2 \)

// Initial Pose Optimization

\( P_0 \leftarrow R_0 P_{init} \)

// Iterative Optimization and Denoising

\( r \leftarrow K^{-1} p_{2d} \)

\( \hat{r} \leftarrow \frac{r}{\|r\|_2} \)

for \( i \leftarrow 0 \) to \( n - 1 \) do

if \( i < \text{warmup} \) then

\( T_i \leftarrow T_0 \)

else

\( T_i \leftarrow \arg \min_{T_i} \left\| C_{2d}(K(P_i + T_i) - p_{2d}) \right\|_2 \)

end if

// Project 3D keypoints to rays

\( \hat{P}_i \leftarrow (P_i + T_i \cdot \hat{r}) \hat{r} - T_i \)

\( P_{i+1} \leftarrow \theta_g(\hat{P}_i, t(i)) \)

end for

return \( P_n \)

3.3. Optimizer in the Loop

In previous works [1, 12, 37, 54], to optimize an accurate 3D human pose, it requires a lot of kinematic constraints, which call for a strong domain knowledge about human motion. Different from those previous works, as shown in Alg 1, our iterative optimizer utilizes the denoising capability and learned human pose prior of the pose generation diffusion model to estimate an accurate 3D human pose with a simple yet effective optimization pipeline without any explicit kinematic constraint.

With a camera intrinsic matrix, \( K \), 2D keypoints, \( p_{2d} \), can be converted to 3D rays, \( r \in \mathbb{R}^{J \times 3} \), based on perspective projection,

\[
r = K^{-1} p_{2d}, \quad \hat{r} = \frac{r}{\|r\|_2}. \quad (3)
\]

Intuitively, as shown in Fig 4, projecting the 3D keypoints from \( P_0 \) to \( r \) will minimize the 2D re-projection error and provide the estimated 3D human poses,

\[
\hat{P}_0 \leftarrow ((P_0 + T) \cdot \hat{r}) \hat{r} - T. \quad (4)
\]

However, there are two problems: 1) Simply projecting 3D keypoints from \( P_0 \) to \( r \) generates a noisy 3D human pose, which may not satisfy the kinematic constraints of the human body. 2) With different translations between \( P_0 \) and the camera, there are different projection results.

In order to solve these two problems, we need to ensure the estimated 3D poses, \( P_1 \), are valid poses and inherently
follow the kinematic constraints to find the optimal translation, \( T_i \). This calls for our use of pose prior.

**Pose prior** Although there is no re-projection error from the optimized poses, the optimized poses may not satisfy the kinematic constraints. Therefore, in previous works, kinematic constraints are added to the pose optimization pipeline, and a complex joint optimization problem is designed to find optimal poses. However, in our pipeline, we take advantage of a pre-trained diffusion-based pose generation model to 'denoise' our optimized poses. As mentioned in DDPM [16], DDIM [48] and Score Matching network [50], the diffusion model is trained by maximizing likelihood, which aims at finding the most possible valid pose based on the input noisy pose. As a result, we use the diffusion-based pose generation model to find the optimal \( P_{i+1} \) based on optimized pose \( \bar{P}_i \).

\[
P_{i+1} = \theta_g(\bar{P}_i, t(i)), \tag{5}
\]

which is different from other diffusion-based methods, like GFPose [8] and DiffPose [10], \( P = \theta(x, c, t) \), where \( x \) is random noise, \( c \) is pose condition and \( t \) is timestamp.

However, during training, the reverse diffusion starts from the standard Gaussian noise, \( \mathcal{N}(0, 1) \), but in our case, the generation model is utilized to denoise an optimized pose, which does not follow the standard Gaussian noise. Inspired by [36, 66], we adopt truncated diffusion model inference, whose timestamp is truncated from the training timestamp during inference. In our case, the timestamp, \( t \), is truncated as \( t \in \{0, 0.1\} \), instead of \( (0, 1] \).

**Find optimal translations** Optimal translations are derived from \( T_o \) by minimizing the 2D re-projection error of \( P_i \), depending on the current iteration number, since the optimized pose, \( P_i \), is not reliable enough in the early iterations.

After certain iterations, the optimal translation is derived from the following:

\[
T_i \leftarrow \arg\min_{T_i} \left\| C_{2d}(K(P_i + T_i) - p_{2d}) \right\|_2^2, \tag{6}
\]

where \( C_{2d} \in \mathbb{R}^J \) are the confidence scores of 2D keypoints, \( p_{2d} \) and \( K \) are the camera intrinsic matrices. Inspired by [18], \( C_{2d} \) are helpful to guide the translation optimization. There is a closed-form solution to the optimization problem. The details can be found in the supplementary material.

By incorporating optimal translations and kinematic constraints, we can optimize the 3D human poses by iteratively projecting \( P_i \) to \( r \) and denoising the projected \( P_{i+1} \) with the help of the diffusion model.

### 4. Experimental Results

In this section, we will introduce the experimental results of ZeDO on 3DPW [57], Human3.6M [17] and MPI-INF-3DHP [34] datasets. More results on Ski-Pose [43] datasets are included in the supplementary materials. Since ZeDO requires an initial pose as a starting point for optimization, and different initial poses may lead to different pose estimation results, we report our results in Mean Per Joint Position Error (MPJPE) with a single initial pose (single-hypothesis) or minimum MPJPE (minMPJPE) with multiple initial poses, in order to have fair comparison with previous multi-hypothesis 3D HPE methods [8, 45, 59].

#### 4.1. Datasets

**Human3.6M** [17] is the most widely used single-person 3D pose benchmark with more than 3.6 million frames and corresponding 3D human poses. The dataset is collected within a 4 m × 3 m indoor environment, with 11 professional actors (6 males and 5 females) performing 17 distinct actions such as discussion, smoking, capturing photographs, posing, greeting, and talking on the phone. Following the convention of previous works [26, 39, 64] for fair comparison,

![Figure 4. By projecting the \( P_i \) to \( r \), we minimize the 2D re-projection error and find an optimized pose \( \bar{P}_i \).](image-url)

<table>
<thead>
<tr>
<th>Methods</th>
<th>CE</th>
<th>Opt</th>
<th>PA-MPJPE ↓</th>
<th>MPJPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolotouros et al. [24]</td>
<td>59.2</td>
<td>96.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kocabas et al. [22]</td>
<td>51.9</td>
<td>82.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kocabas et al. [23]</td>
<td>46.4</td>
<td>74.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. [25]</td>
<td>45.0</td>
<td>74.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ma et al. [32]</td>
<td>41.3</td>
<td>67.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. [25]</td>
<td></td>
<td>50.9</td>
<td>82.0</td>
<td></td>
</tr>
<tr>
<td>Kocabas et al. [22]</td>
<td></td>
<td>56.5</td>
<td>93.5</td>
<td></td>
</tr>
<tr>
<td>Kocabas et al. [23]</td>
<td></td>
<td>50.9</td>
<td>82.0</td>
<td></td>
</tr>
<tr>
<td>Gong et al. [11]</td>
<td></td>
<td>58.5</td>
<td>94.1</td>
<td></td>
</tr>
<tr>
<td>Gholami et al. [9]</td>
<td></td>
<td>46.5</td>
<td>81.2</td>
<td></td>
</tr>
<tr>
<td>Chai et al. [5]</td>
<td></td>
<td>55.3</td>
<td>87.7</td>
<td></td>
</tr>
<tr>
<td>Song et al. [47]</td>
<td></td>
<td>55.9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Choutas et al. [6]</td>
<td></td>
<td>52.2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>ZeDO</strong> (( S = 1, J = 17 ))</td>
<td>✓</td>
<td>✓</td>
<td><strong>40.3</strong></td>
<td><strong>69.7</strong></td>
</tr>
<tr>
<td><strong>ZeDO</strong> (( S = 1, J = 14 ))</td>
<td>✓</td>
<td>✓</td>
<td><strong>43.1</strong></td>
<td><strong>76.6</strong></td>
</tr>
</tbody>
</table>

| Table 1. Cross-domain evaluation results on 3DPW dataset. CE stands for cross-domain evaluation, and Opt means optimization-based method. Ground truth 2D poses are used. |
The learning schedule comes with warmup in the first 5k iterations.

The 3DPW frame test dataset.

Following 60 video sequences captured in the dataset with 18 different poses and dynamic movements. Following actions like walking and sitting to more complex exercises, each encompassing a range of activities from simple to complex.

Consisting of 4 males and 4 females with 8 different poses and dynamic movements. The 3DHP dataset captures poses of 8 actors, outdoors.

The 3DHP dataset captures poses of 8 actors, consisting of 4 males and 4 females with 8 different actions, each encompassing a range of activities from simple actions like walking and sitting to more complex exercise poses and dynamic movements. Following [11], we use a sampled 2929 frame test dataset.

3DPW [57] is the first dataset in-the-wild with accurate 3D poses for evaluation. Compared with Human3.6M and MPI-INF-3DHP, 3DPW focuses on outdoor scenarios and captures videos with static and moving cameras. There are 60 video sequences captured in the dataset with 18 different actors. Following [5], we test ZeDO on 3DPW only for cross-dataset evaluation.

### 4.2. Training and Inference Details

We pre-train our 3D pose generation model for 5000 epochs on one NVIDIA A100 with a batch size of 50k, a learning rate of $2e^{-4}$ with an Adam optimizer. The training schedule comes with warmup in the first 5k iterations and cosine learning rate decay in the following iterations.

As [50], the timestamp, $t$, during the forward or reverse diffusion process, is uniformly sampled from $(0, 1)$. All 3D human poses are normalized to pelvis-related coordinates during training and inference. To improve the robustness of the model, we apply flip and rotation data augmentation during training. For cross-domain evaluation, we pre-train the pose generation model in a different dataset and directly test the optimization pipeline without any fine-tuning.

During inference, the pipeline supports single or multiple initial poses. For the initial pose optimizer, we limit the rotation axis to the z-axis only for better performance. We set the number of warmup iterations as 200, and the number of total iterations as 1000. The timestamp, $t$, is uniformly sampled from $(0, 0.1)$. The initial poses are sampled from training sets of Human3.6M [17] or 3DHP [34] by the K-Means algorithm. For different numbers of hypotheses, we run K-Means with different numbers of clusters.

### 4.3. Results

#### 3DPW

3DPW is a challenging in-the-wild dataset compared with Human3.6M and MPI-INF-3DHP datasets. 3DPW focuses on outdoor scenarios with both static and moving cameras. For cross-domain evaluation on the 3DPW dataset, we pre-train the pose generation model on the Human3.6M dataset and inference on the 3DPW dataset without any fine-tuning. On 3DPW, we find some of the previous works [23, 25] evaluate on 14 Leeds Sports Pose (LSP) [20] keypoints, while others [5, 9, 11] evaluate on 17 Human3.6M keypoints, and some other works [24] do not explain clearly which keypoints they use. In Table 1, we report the results of both 14 and 17 keypoints for a fair comparison. We achieve SOTA performance, PA-MPJPE 40.3mm. with a single hypothesis.

#### Human3.6M

Although ZeDO is a Zero-shot Diffusion-based Optimization pipeline, ZeDO achieves...
Table 4. The ablation study results of ZeDO. RO stands for rotation optimization as the initial pose optimization. WU denotes the warmup iterations. RA is the rotation data augmentation for training the pose generation model. The dataset name under the Dataset column is the testing dataset, and the name under the Diff Model column is the dataset used for diffusion model pre-training.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Diff Model</th>
<th>RO</th>
<th>WU</th>
<th>RA</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>H36M</td>
<td>H36M</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3DHP</td>
<td>3DHP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 5. Different diffusion backbone 3D HPE quantitative results on 3DPW dataset.

<table>
<thead>
<tr>
<th>Diffusion Backbone</th>
<th>[PA-MPJPE ↓]</th>
<th>[MPJPE ↓]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score Matching [50]</td>
<td>40.3</td>
<td>69.7</td>
</tr>
<tr>
<td>DDIM [48]</td>
<td>40.4</td>
<td>67.9</td>
</tr>
<tr>
<td>DDPM [38]</td>
<td>51.7</td>
<td>81.3</td>
</tr>
</tbody>
</table>

Table 6. The pose generation models trained on mixed datasets (Human3.6M + 3DHP) achieves the best performance with $S = 50$, as well as better generalization. GT 2D keypoints are used.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Diff Model</th>
<th>MPJPE ↓</th>
<th>PA-MPJPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>H36M</td>
<td>H36M</td>
<td>37.0</td>
<td>27.5</td>
</tr>
<tr>
<td>H36M</td>
<td>mixed</td>
<td>35.7</td>
<td>26.5</td>
</tr>
<tr>
<td>3DHP</td>
<td>H36M</td>
<td>69.9</td>
<td>49.0</td>
</tr>
<tr>
<td>3DHP</td>
<td>3DHP</td>
<td>55.2</td>
<td>38.6</td>
</tr>
<tr>
<td>3DHP</td>
<td>mixed</td>
<td>52.4</td>
<td>37.7</td>
</tr>
</tbody>
</table>
tion. However, the 3D pose distributions vary across different datasets. To ensure the pre-trained diffusion model can be adapted to different datasets, we utilize rotation and flip data augmentation during training. As expected, the data augmentation significantly improves the performance in cross-domain evaluation by 13.9 mm in MPJPE, shown in Table 4.

**Boost the performance further by mixing dataset.** According to Table 6, the pose generation model trained on the mixed datasets (Human3.6M + 3DHP) improves the performance of ZeDO by 1.3mm on Human3.6M and 2.8mm on 3DHP while using the same estimation algorithm.

**How to pick the initial pose?** We utilize K-Means in our experiments to select anchor poses from the training set as initial poses. K-Means effectively finds the most representative poses in the training set, making it superior to other sampling strategies, as shown in Table 7. Random Sampling randomly samples a pose from the training set, whereas Random Generation is generated by the pre-trained pose generation model.

**What is the best number of optimization iterations?** In ZeDO, the number of diffusion optimization iterations is set to 1000. Intuitively, increasing the number of iterations can enhance performance but may suffer the inference speed. In Fig 5, as expected, the inference time increases linearly with respect to the number of iterations. However, the figure shows that the best performance is achieved when the number of iterations is around 1000. With the number of iterations exceeding 1000, there is no performance gain, and the inference speed decreases.

### 5.2. Limitations

Although ZeDO achieves state-of-the-art performance in various benchmarks and settings, several limitations still need to be further explored. 1) Similar to other optimization-based approaches, the optimizer in the loop requires camera intrinsic parameters. 2) Since our method is based on minimizing the 2D re-projection error, we are not able to solve the ambiguity of the depth and scale without additional information like bone length or height. 3) For an identical 2D human pose, there are multiple 3D human poses matched. Without image or temporal information, the 1-to-many mapping issue cannot be resolved by single frame lifting methods, as shown in Fig 6.

### 6. Conclusion

In this paper, we propose ZeDO, a Zero-shot Diffusion-based Optimization pipeline for 3D HPE. To the best of our knowledge, we are the first to introduce the diffusion model to the optimization-based method in the 3D HPE task. We leverage the pre-trained diffusion-based 3D human pose generation model and can optimize target 3D poses in the loop. To be specific, an optimizer that calculates the optimal translation is used iteratively with denoising steps in the diffusion model. Compared to other prior arts, ZeDO achieves state-of-the-art performance on Human3.6M, MPI-INF-3DHP, and 3DPW datasets, even with cross-dataset evaluation. In the future, we plan to further improve ZeDO by modifying the diffusion model and solving the limitations listed in Sec 5. It is our wish that this optimization method could become a common paradigm beyond end-to-end lifting networks in 3D human pose estimation tasks.
References


[26] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d


