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

Zhongyu Jiang[1], Zhuoran Zhou[1], Lei Li[2], Wenhao Chai[1], Cheng-Yen Yang[1], Jenq-Neng Hwang[1]
[1]University of Washington, [2]University of Copenhagen

Motivation

Most 2D-3D human pose estimation models are learning-based and are susceptible to domain gaps, leading to low generality across new datasets. However, optimization methods suffer from low performance but can function case by case. Our method takes advantage of both by proposing an optimization algorithm that applies a 3D prior diffusion model to predict accurate 3D poses based on 2D keypoints. Additionally, our method shows great zero-shot and cross-evaluation capacities across 3DPW, H36m and 3DHP.

Method

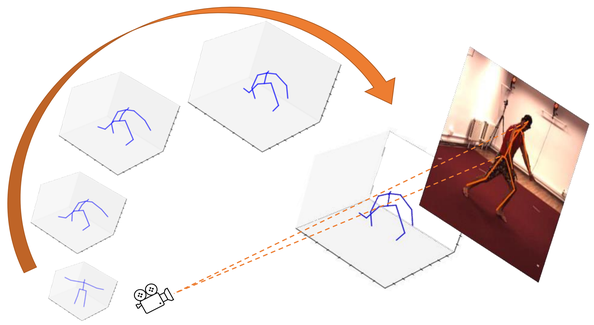
Algorithm 1 ZeDO pipeline

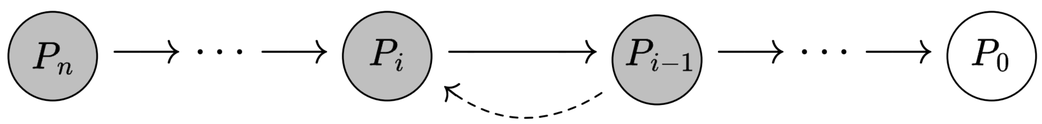
Require: Initial 3D pose P_{init}, Target 2D pose p_{2d}, 2D pose confidence scores C_{2d}, Camera intrinsic K, Diffusion timestamp t, Pre-trained diffusion model $\theta_g(P, t)$

$R_0, T_0 \leftarrow \arg\min_{R_0,T_0} \|K(R_0 P_{init} + T_0) - p_{2d}\|_2$
// 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 $i < warmup$ **then**
    $T_i \leftarrow T_0$
  else
    $T_i \leftarrow \arg\min_{T_i} \|C_{2d}(K(P_i + T_i) - p_{2d})\|_2$
  end if
  // Project 3D keypoints to rays
  $\widetilde{P}_i \leftarrow \left((P_i + T_i) \cdot \hat{r}\right)\hat{r} - T_i$
  $P_{i+1} \leftarrow \theta_g(\widetilde{P}_i, t(i))$
end for
return P_n

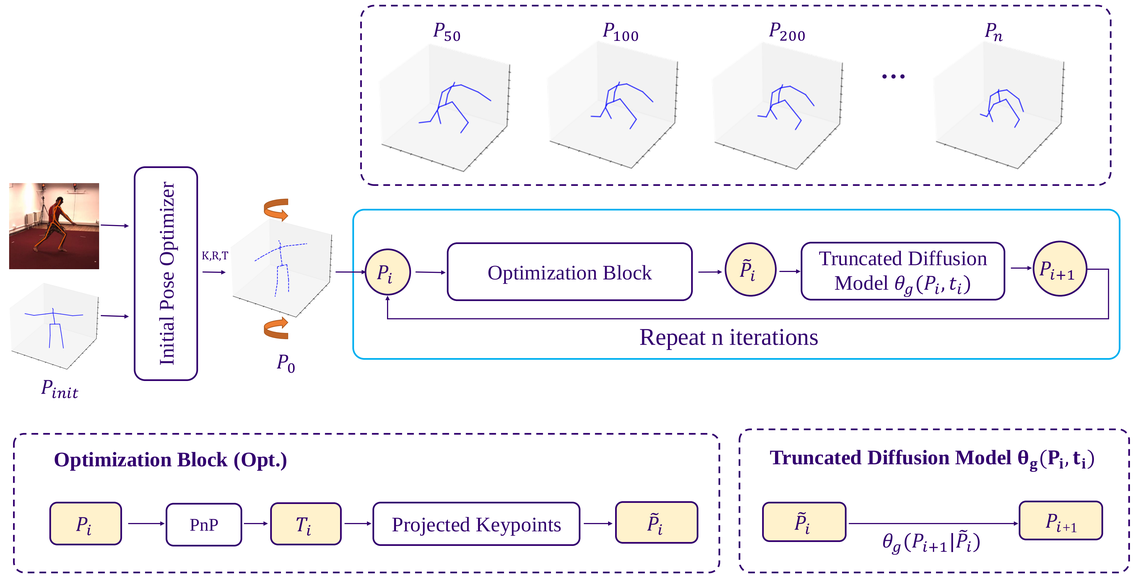
Algorithm Outline :

1. Set KNN-cluster as the initial pose. Run optimization for appropriate R and T.
2. Compute camera rays with camera parameters and 2D pose.
3. Move 3D pose to the rays to minimize projection error. Update the new T unless it is during warmup.
4. Adjust 3D pose by the prior diffusion model.
5. Repeat steps 2-5 1000 times.

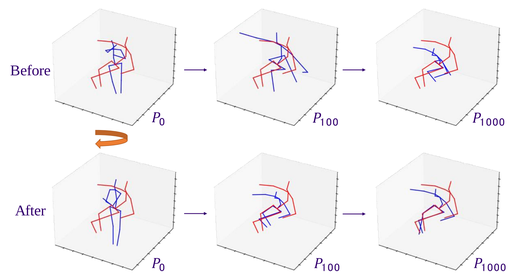




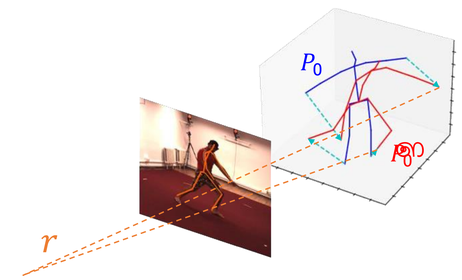
We apply diffusion model as your 3D Human Pose prior generator.



Our method's pipeline consists of four components illustrated in the algorithm outline. The optimization block and truncated diffusion model repeat 1000 iterations to achieve an accurate 3D prediction, P_{1000}.



Initial Optimization



Minimize Reprojection Error

Experiment

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

Results on 3DPW

Dataset	Diff Model	RO	WU	RA	GT	$S = 1$ MPJPE ↓	$S = 1$ PA-MPJPE ↓	$S = 50$ MPJPE ↓	$S = 50$ PA-MPJPE ↓
H36M	H36M					75.0	52.7	53.4	42.7
H36M	H36M	✓				77.2	53.7	52.7	42.4
H36M	H36M	✓	✓			**65.7** (9.3 ↓)	**49.0** (3.7 ↓)	**51.4** (2.0 ↓)	**42.1** (0.6 ↓)
H36M	H36M	✓	✓	✓		69.5	51.4	52.9	42.5
H36M	H36M	✓	✓		✓	50.1	35.8	37.0	27.5
3DHP	H36M				✓	148.3	88.8	93.4	59.0
3DHP	H36M	✓	✓		✓	113.8	74.1	80.1	56.0
3DHP	H36M	✓	✓	✓	✓	**99.9** (48.4 ↓)	**67.9** (20.9 ↓)	**69.9** (23.5 ↓)	**49.0** (10.0 ↓)
3DHP	3DHP	✓	✓		✓	86.5	55.9	55.2	38.6

Ablation Study on H36M and 3DHP

Code