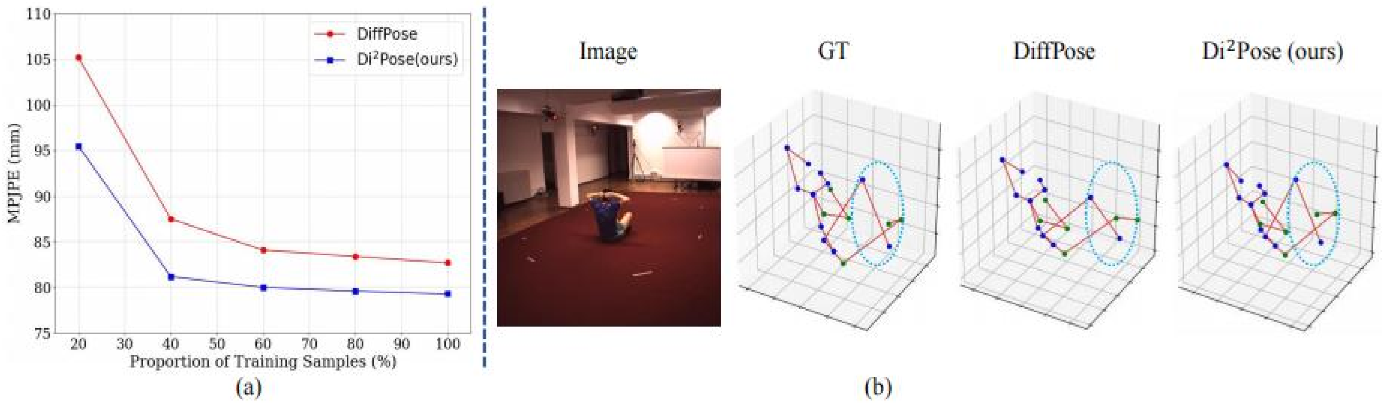
Di²Pose: Discrete Diffusion Model for Occluded 3D Human Pose Estimation

Weiquan Wang[1] Jun Xiao[1] Chunping Wang[2] Wei Liu[3] Zhao Wang[1] Long Chen[4*]

[1]Zhejiang University [2]Finvolution Group [3]Tencent [4]HKUST

Motivation



(a) Results of DiffPose and Di²Pose in Human3.6M dataset (with MPJPE metric), across varying proportions of training samples
(b) Prediction results of two methods under occlusion

- **Challenge in 3D HPE**: Accurately estimating 3D human poses from monocular images is difficult, especially when *occlusions cause uncertainty and ambiguity.*
- **Limitations of Existing Methods**: Mainstream approaches *overlook interdependencies between joints* by treating them independently, leading to inaccuracies under occlusions.
- **Scarcity of 3D Pose Data**: Diffusion-based models need large datasets, but *limited 3D pose data* can result in implausible poses that don't reflect human biomechanics, especially in occluded scenarios.

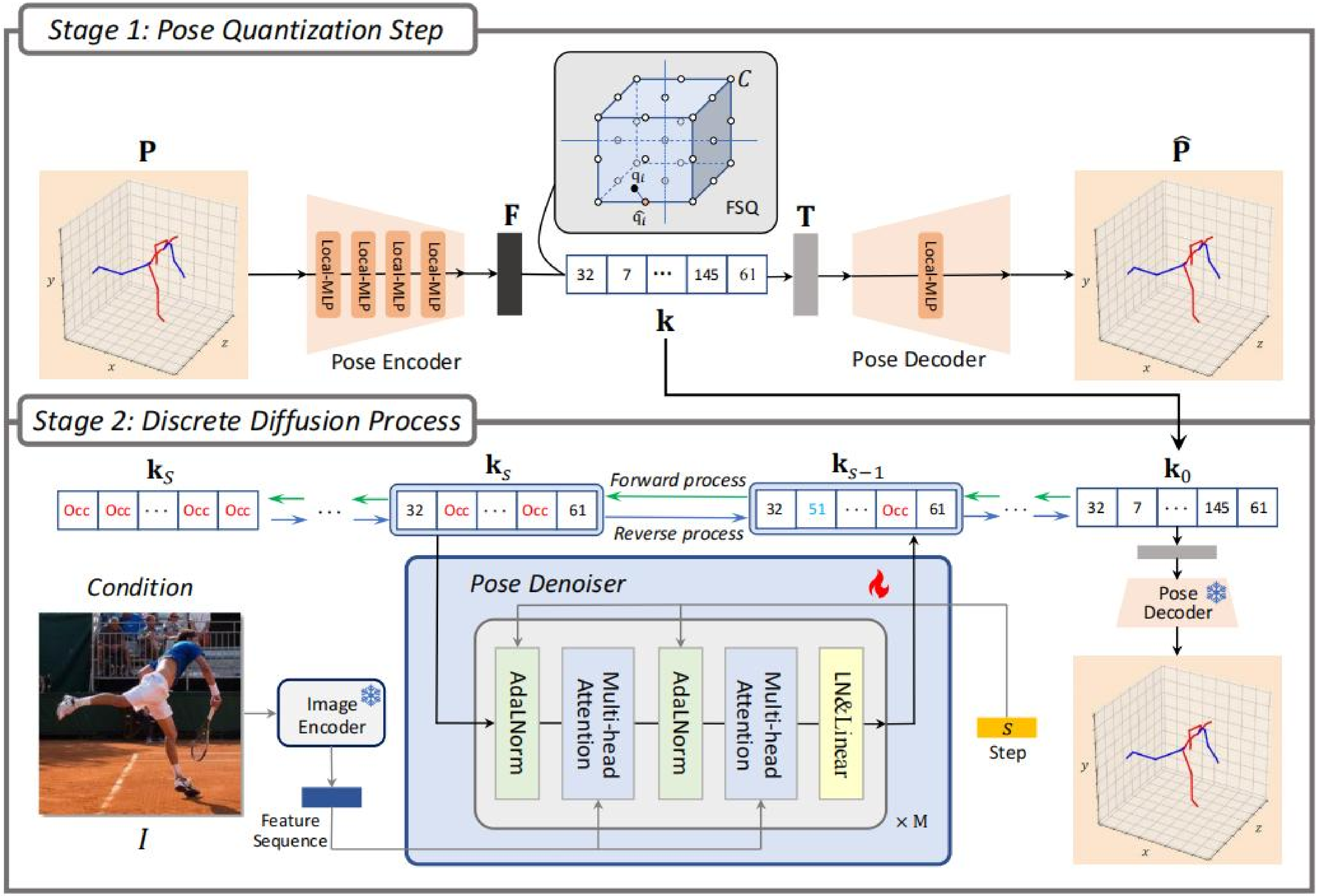
Contribution

- **Di²Pose framework** integrates the inherent discreteness of 3D pose data into the diffusion model, offering *a new paradigm for addressing 3D HPE under occlusions*.

- The designed **pose quantization step** represents 3D poses in a compositional manner, effectively *capturing local correlations between joints* and confining search space to reasonable configurations.

- The constructed **discrete diffusion process** *simulates the complete process of a 3D pose transitioning from occluded to recovered*, which introduces the impact of occlusions into pose estimation process.

Method


Stage 1: Pose Quantization Step
P
Local-MLP
Local-MLP
Local-MLP
Local-MLP
Pose Encoder
F
C
FSQ
32
7
···
145
61
k
T
Local-MLP
Pose Decoder
P̂
Stage 2: Discrete Diffusion Process
$\mathbf{k}_S$
Occ
Occ
···
Occ
Occ
$\mathbf{k}_s$
32
Occ
···
Occ
61
Forward process
Reverse process
$\mathbf{k}_{s-1}$
32
51
···
Occ
61
$\mathbf{k}_0$
32
7
···
145
61
Condition
I
Image Encoder
Feature Sequence
Pose Denoiser
AdaLNorm
Multi-head Attention
AdaLNorm
Multi-head Attention
LN&Linear
× M
s
Step
Pose Decoder

Experimental Results

Quantitive Results

Human3.6M

Table 1: Results on Human3.6M in millimeters under MPJPE. The best results are in **bold**, and the second-best ones are underlined.

Methods	Dir	Disc	Eat	Gr.	Phon.	Phot.	Pose	Pur.	Sit	SitD.	Sm.	Wait	W.D.	Walk	W.T.	Avg
Pavlakos *et al.* [54] *CVPR'17*	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 *et al.* [48] *ICCV'17*	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
Hossain *et al.* [29] *ECCV'18*	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
Zhao *et al.* [85] *CVPR'19*	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
Liu *et al.* [45] *ECCV'18*	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 *et al.* [77] *CVPR'21*	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 *et al.* [88] *CVPR'22*	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
Geng *et al.* [21] *CVPR'23*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50.8
Choi *et al.* [14] *IROS'23*	44.3	51.6	46.3	51.1	**50.3**	54.3	49.4	45.9	57.7	71.6	**48.6**	49.1	52.1	44.0	44.4	50.7
Zhang *et al.* [82] *TPAMI'23*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50.2
Gong *et al.* [22] *CVPR'23*	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
Di^{2}Pose (Ours)	**41.9**	**47.8**	**45.0**	49.0	51.5	62.2	**45.7**	45.6	**57.6**	67.1	50.1	**45.3**	**51.4**	**37.3**	40.9	**49.2**

3DPW

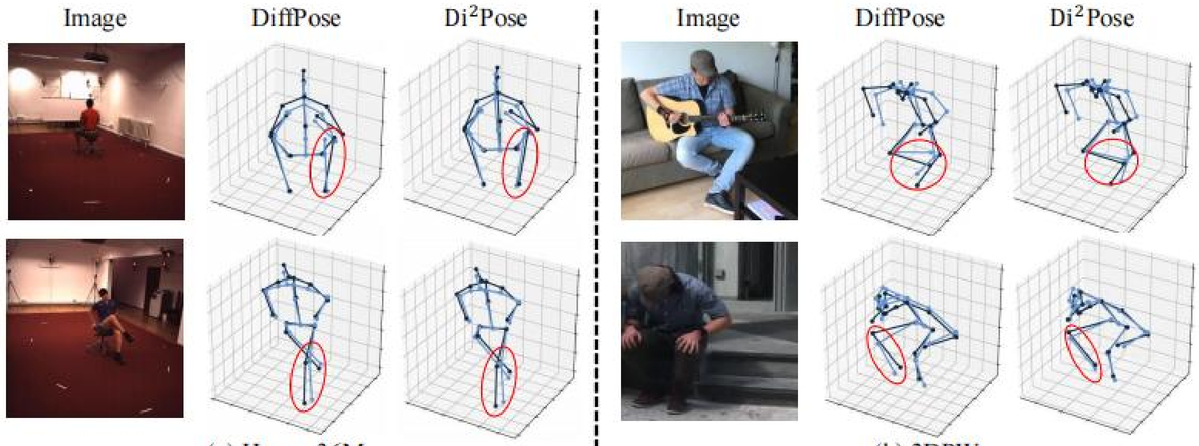
Table 2: Evaluation on 3DPW, 3DPW-Occ, and 3DPW-AdvOcc. The number 40 and 80 after 3DPW-AdvOcc denote the occluder size. * denotes the results from our implementation. The best results are in **bold**, and the second-best ones are underlined.

Methods	3DPW [72]		3DPW-Occ [83]		3DPW-AdvOcc@40		3DPW-AdvOcc@80	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Cai *et al.* [9] *ICCV'19*	112.9	69.6	115.8	72.3	241.1	101.4	355.9	116.3
Pavllo *et al.* [55] *CVPR'19*	101.8	63.0	106.7	67.1	221.6	99.4	334.3	112.9
Cheng *et al.* [12] *AAAI'21*	—	64.2	—	85.7	279.4	113.2	371.4	119.8
Zheng *et al.* [90] *ICCV'21*	118.2	73.1	132.8	80.5	247.9	106.2	359.6	115.5
Zhang *et al.* [82] *TPAMI'23*	91.1	54.3	94.6	56.7	142.5	73.8	251.8	103.9
Geng *et al.* * [21] *CVPR'23*	83.1	53.9	82.8	53.7	127.2	71.9	192.5	92.1
Gong *et al.* * [22] *CVPR'23*	82.7	53.8	82.1	53.5	121.4	70.9	189.3	92.4
Di^{2}Pose (Ours)	**79.3**	**50.1**	**79.6**	**50.7**	**108.4**	**59.8**	**153.6**	**78.7**

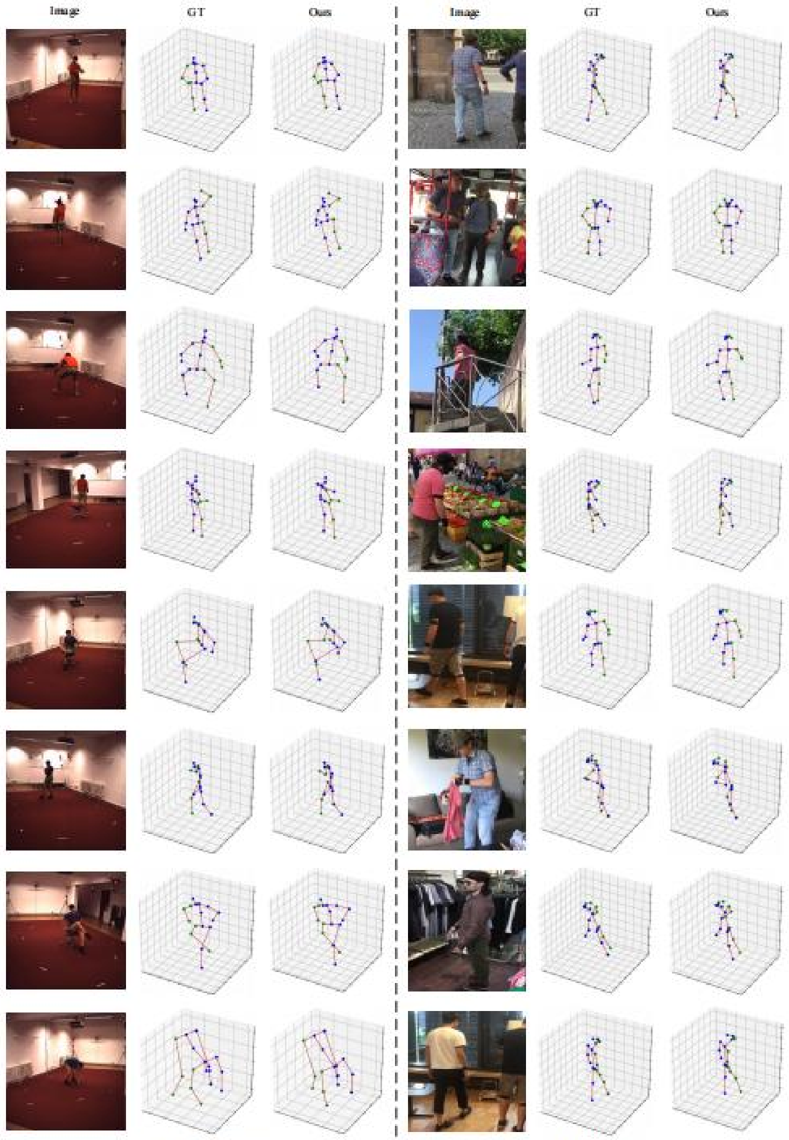
Experimental Results

NEURAL INFORMATION PROCESSING SYSTEMS

Qualitive Results



(a) Human36M (b) 3DPW



(a) Human3.6M (b) 3DPW

Conclusion

- *We presents Di²Pose, a novel diffusion-based framework that tackles occluded 3D HPE in discrete space*
- *Di²Pose first captures the local interactions of joints and represents a 3D pose by multiple quantized tokens. Then, the discrete diffusion process models the discrete tokens in latent space through a conditional diffusion model, which implicitly introduces occlusion into the modeling process for more reliable 3D HPE with occlusions*
- *Experimental results show that our method surpasses the state-of-the-art approaches on three widely used benchmarks*

Future Work

Extend Di²Pose to video datasets to leverage interframe information for enhanced temporal consistency

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for your listening