香港浸會大學 HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系

ICASSP 2024 KOREA

Exploring Latent Cross-Channel Embedding for Accurate 3D Human Pose Reconstruction in a Diffusion Framework

Junkun Jiang
Department of Computer Science,
Hong Kong Baptist University
Hong Kong SAR, China
csjkjiang@comp.hkbu.edu.hk

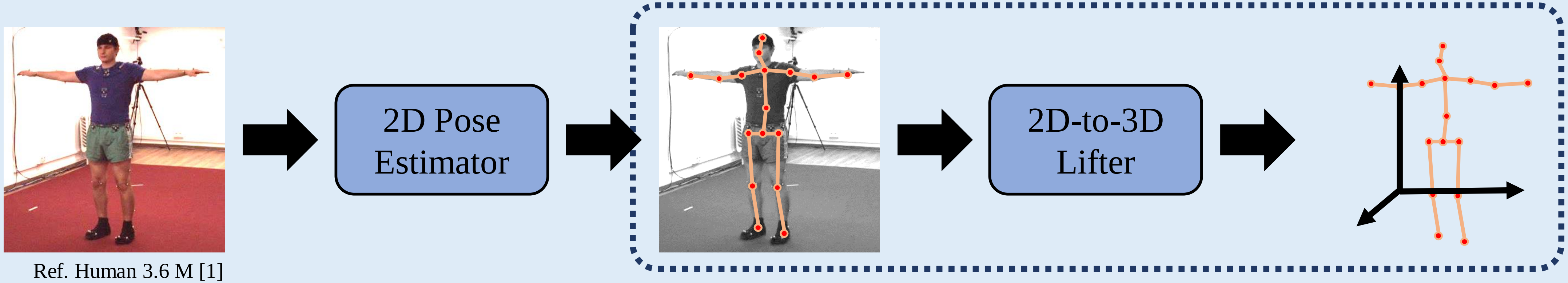
Jie Chen*
Department of Computer Science,
Hong Kong Baptist University
Hong Kong SAR, China
chenjie@comp.hkbu.edu.hk

Speaker: JIANG, Junkun
csjkjiang@comp.hkbu.edu.hk

2024/04/18

Introduction

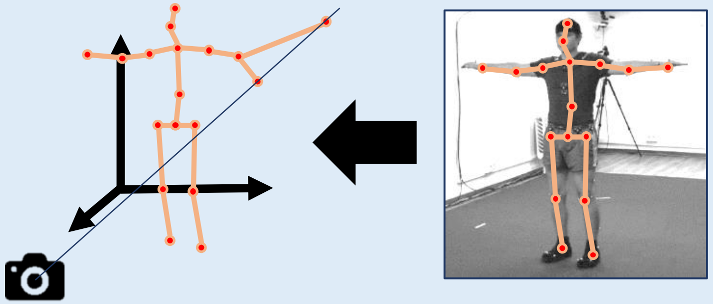
Objective of *Single-frame Monocular 3D Human Pose Estimation*



Ref. Human 3.6 M [1]

Challenges

- Depth ambiguities. Estimating a 3D pose from an image is ill-posed.
- Noisy 2D estimation results, leading to inaccurate 3D lifting.
- Potential occlusions.



Ill-posed problem: One 2D reprojection has multiple 3D poses.

[1] Ionescu, Catalin, et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence 36.7 (2013): 1325-1339.

Introduction

Motivation

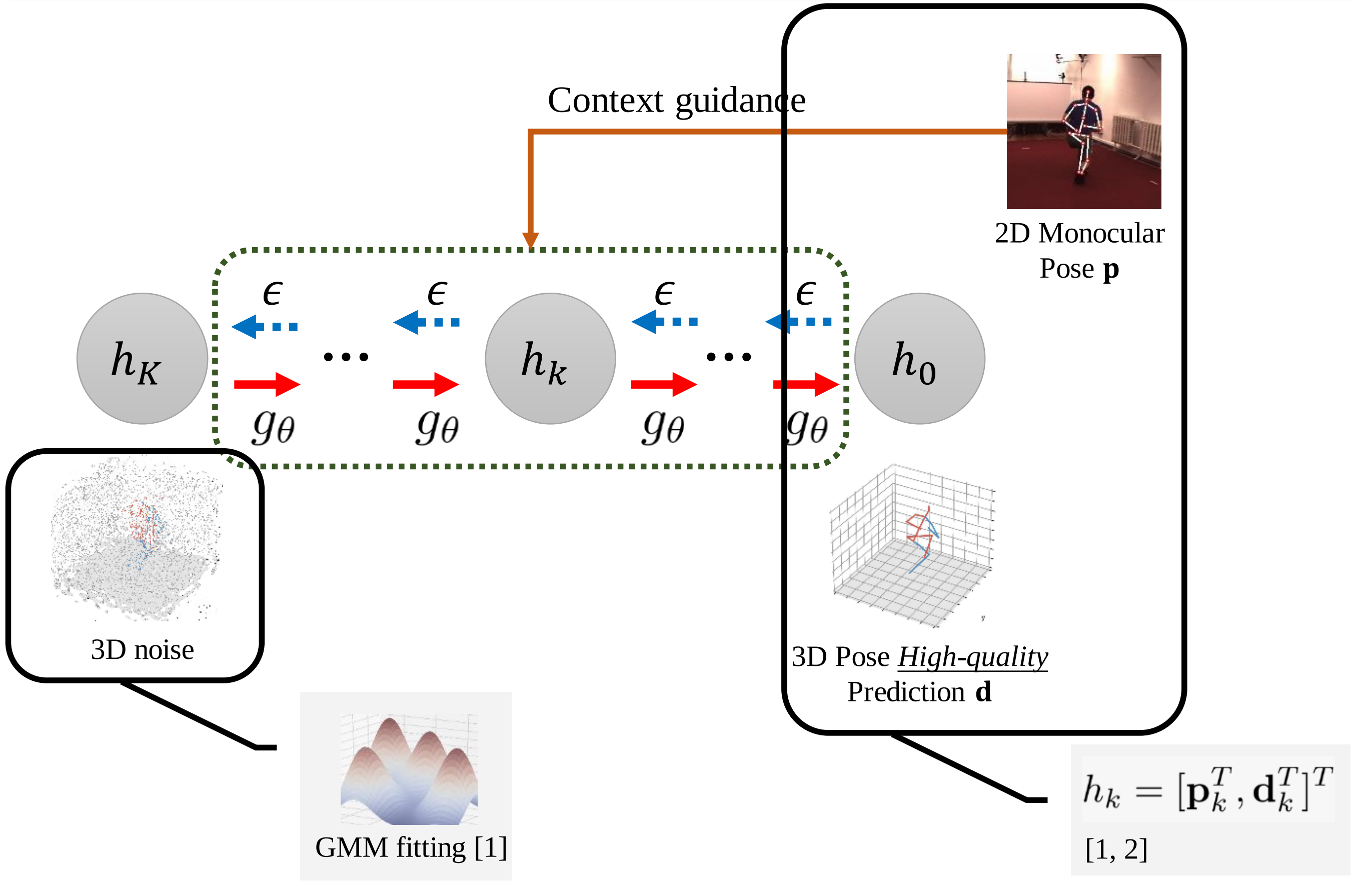
- Existing approaches struggle in over-fitting a projection matrix but not effectively address the above challenges.
- Recent advancements in diffusion models have shown promise in incorporating structural priors to address projection ambiguities.
- However, these methods overlook the exploration of correlation between the 2D and 3D joint level features.

Contribution

- We develop a novel monocular single-frame 3D HPE framework driven by the diffusion model that works as a post-processing to refine the lifting results.
- To explore the correlation between the 2D and 3D joint level features, the Cross-Channel Embedding (CCE) module is proposed inside the framework.
- To encourage efficient cross-joint attention propagation, the Context Attention Guidance (CAG) module is proposed inside the framework.

Methodology

Preliminary



Context guidance

2D Monocular Pose **p**

h_K h_k h_0

ϵ g_θ … ϵ g_θ ϵ g_θ … ϵ g_θ

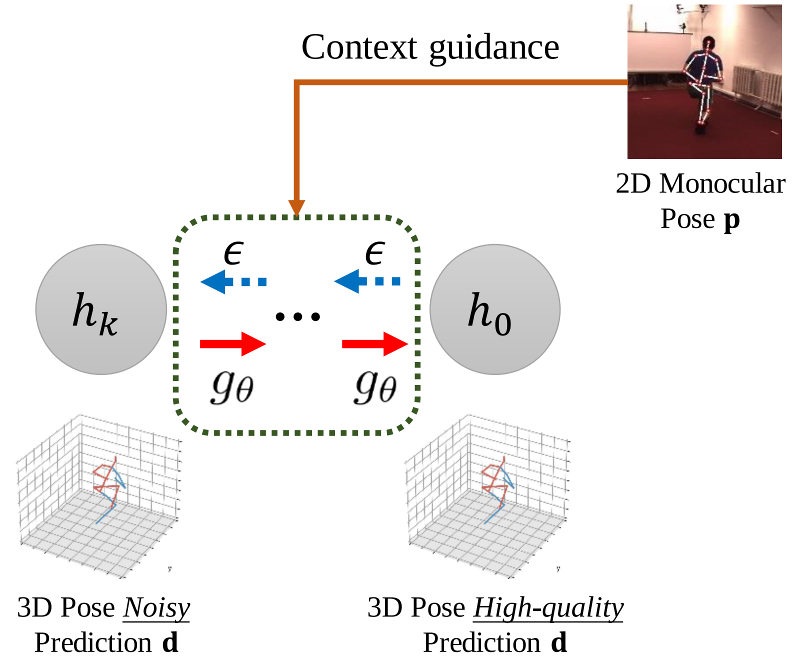
3D Pose *High-quality* Prediction **d**

3D noise

GMM fitting [1]

$$h_k = [\mathbf{p}_k^T, \mathbf{d}_k^T]^T$$

[1, 2]



Context guidance

2D Monocular Pose **p**

h_k h_0

ϵ g_θ … ϵ g_θ

3D Pose *Noisy* Prediction **d**

3D Pose *High-quality* Prediction **d**

To get a pose-specific distribution instead of directly using a normal distribution

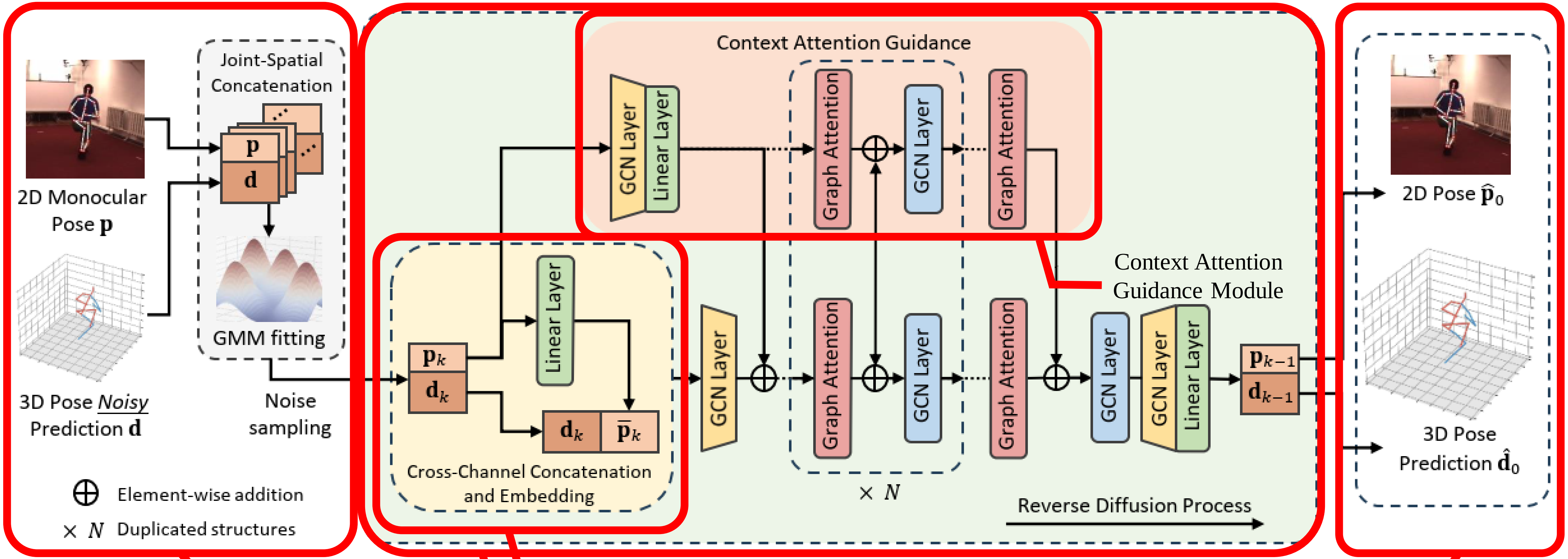
csjkjiang@comp.hkbu.edu.hk

[1] Gong, Jia, et al. "Diffpose: Toward more reliable 3d pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Choi, Jeongjun, Dongseok Shim, and H. Jin Kim. "Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model." 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023.

Methodology

Proposed framework



Context Attention Guidance Module

Noise sampled from GMM

Reverse Diffusion Process

Cross-Channel Embedding Module

High-quality prediction

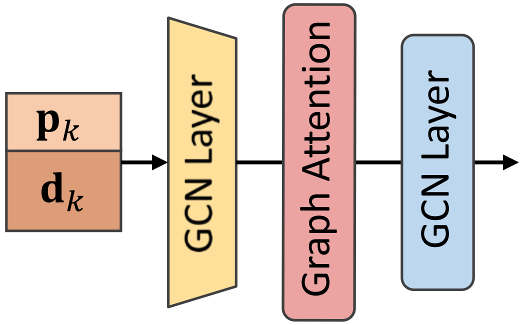
Methodology

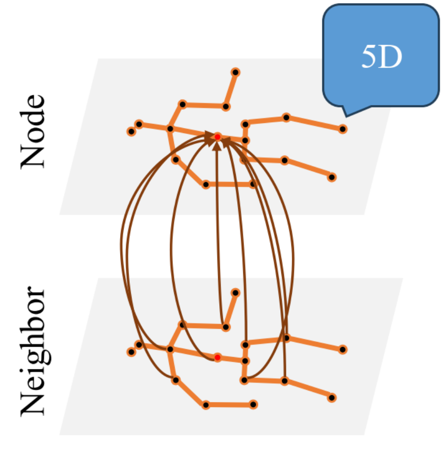
Proposed framework

- Cross-Channel Embedding Module (CCE)

Recent works [1, 2] directly concatenate 2D and 3D joint coordinates as h_k and process h_k with GCN layers.

- *i.e.*, $h_k = [\mathbf{p}_k^T, \mathbf{d}_k^T]^T \in \mathrm{R}^{5\times J}$.
- Each (**p**, **d**) concatenation is regarded as one GCN node.
- Only the cross-joint correlations are explored, leaving the implicit relations between the 2D and lifted 3D features within and across different joints under-investigated.





a) Cross-joint connection

[1] Gong, Jia, et al. "Diffpose: Toward more reliable 3d pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Choi, Jeongjun, Dongseok Shim, and H. Jin Kim. "Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model." 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023.

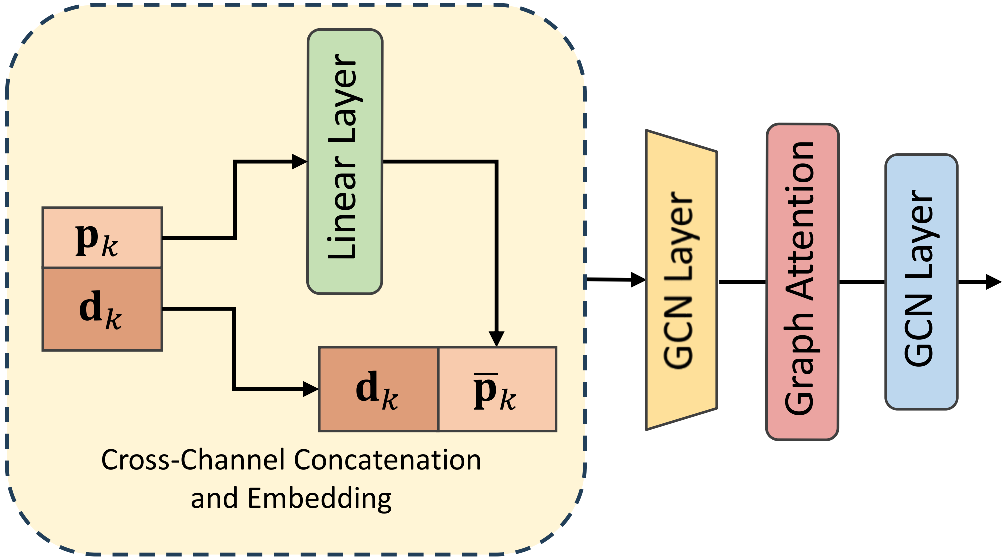
Methodology

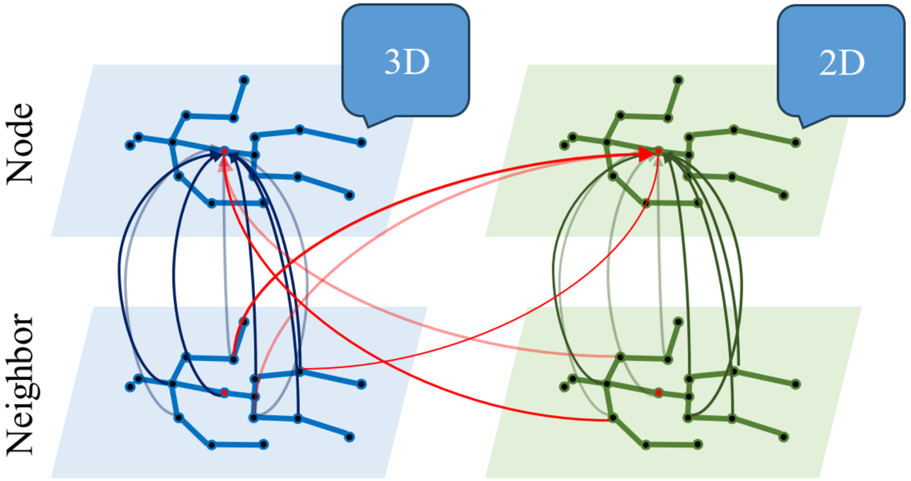
Proposed framework

- Cross-Channel Embedding Module (CCE)

Instead of direct spatial concatenation, we concatenate the 2D and 3D features along their channel dimensions: $h_k = [L_e(\mathbf{p}_k), \mathbf{d}_k] \in \mathrm{R}^{3\times 2J}$, where L_e is a linear projection layer for dimension alignment.

- The connections between 2D and 3D features are comprehensively considered.



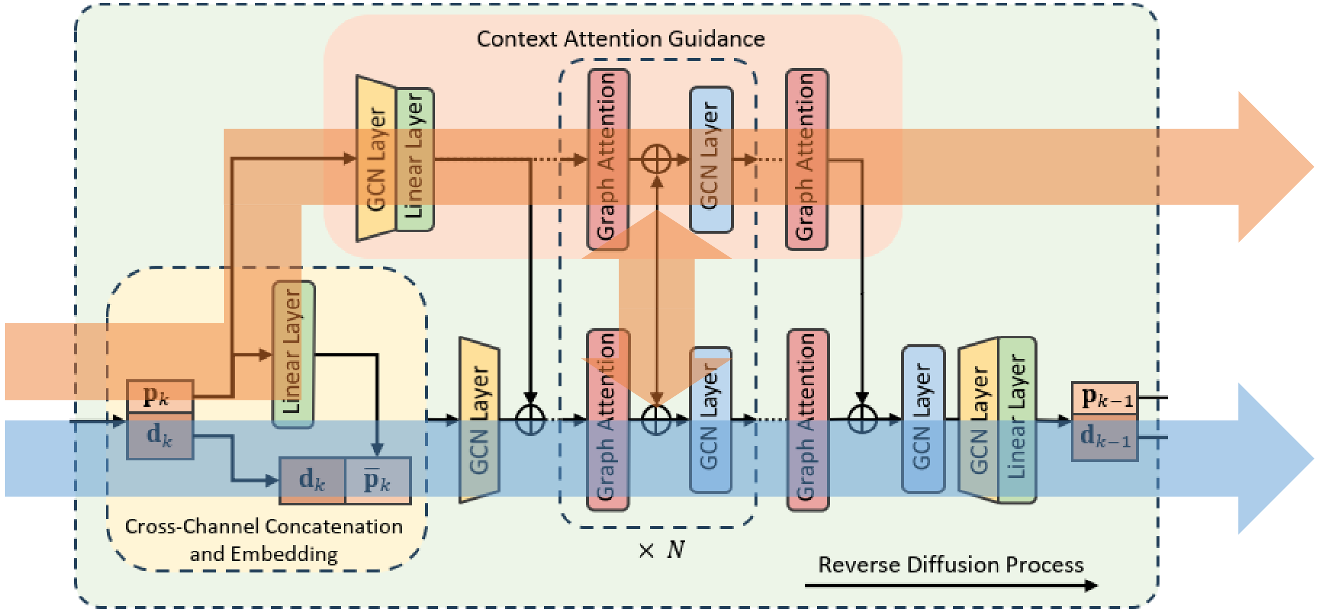


b) Cross-channel connection

Methodology

Proposed framework

- Context Attention Guidance Module (CAG)
  - 2D pose features are first processed by GCN layers.
  - Features of two branches are merged by element-wise addition after self-attention layers.
  - To further encourage efficient cross-joint attention propagation among the latent channels throughout the iterative diffusion process.



Context Attention Guidance branch

Signal diffusion branch

Experiment results

We report quantitative and qualitative results on *Human 3.6 M* and *MPI-INP-3DHP* datasets with the respective input ground-truth and estimated 2D pose. Ablation studies are reported in Row 6 to Row 7.

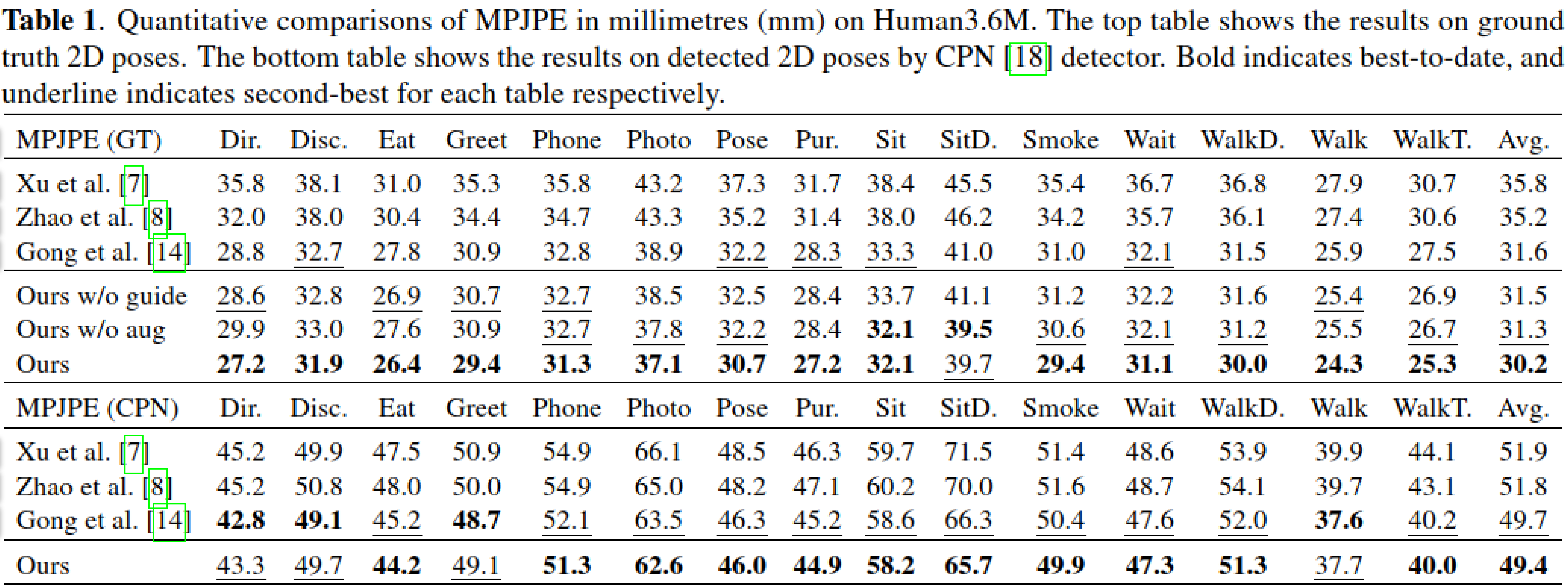


Table 1. Quantitative comparisons of MPJPE in millimetres (mm) on Human3.6M. The top table shows the results on ground truth 2D poses. The bottom table shows the results on detected 2D poses by CPN [18] detector. Bold indicates best-to-date, and underline indicates second-best for each table respectively.

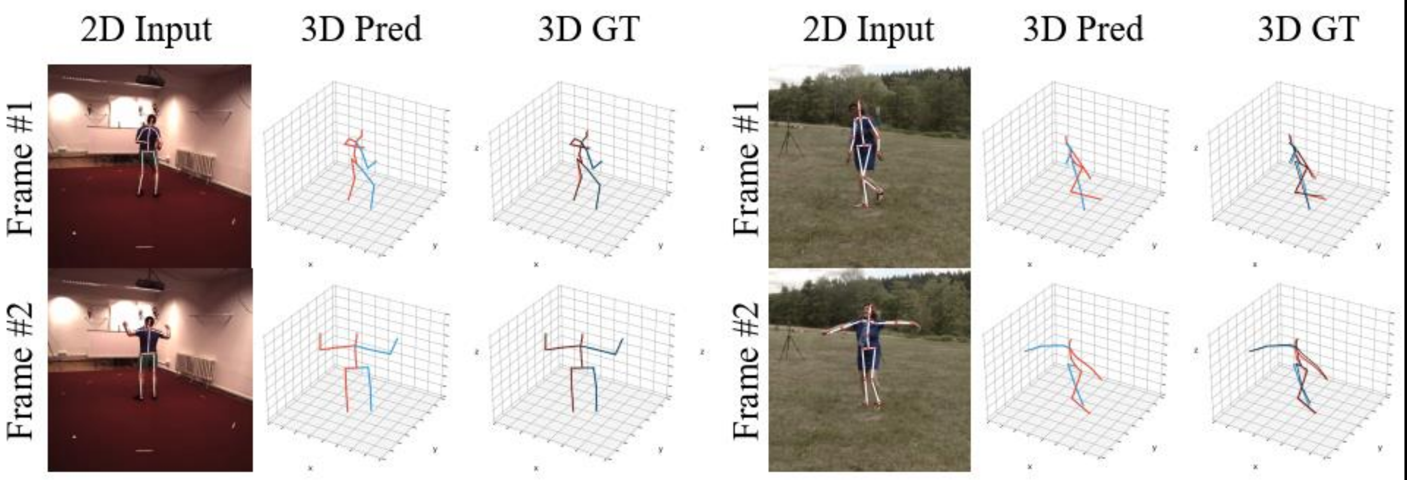
Venue	MPJPE (GT)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
CVPR'21	Xu et al. [7]	35.8	38.1	31.0	35.3	35.8	43.2	37.3	31.7	38.4	45.5	35.4	36.7	36.8	27.9	30.7	35.8
CVPR'22	Zhao et al. [8]	32.0	38.0	30.4	34.4	34.7	43.3	35.2	31.4	38.0	46.2	34.2	35.7	36.1	27.4	30.6	35.2
CVPR'23	Gong et al. [14]	28.8	32.7	27.8	30.9	32.8	38.9	32.2	28.3	33.3	41.0	31.0	32.1	31.5	25.9	27.5	31.6
	Ours w/o guide	28.6	32.8	26.9	30.7	32.7	38.5	32.5	28.4	33.7	41.1	31.2	32.2	31.6	25.4	26.9	31.5
	Ours w/o aug	29.9	33.0	27.6	30.9	32.7	37.8	32.2	28.4	**32.1**	**39.5**	30.6	32.1	31.2	25.5	26.7	31.3
	Ours	**27.2**	**31.9**	**26.4**	**29.4**	**31.3**	**37.1**	**30.7**	**27.2**	**32.1**	39.7	**29.4**	**31.1**	**30.0**	**24.3**	**25.3**	**30.2**
	MPJPE (CPN)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
CVPR'21	Xu et al. [7]	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
CVPR'22	Zhao et al. [8]	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
CVPR'23	Gong et al. [14]	**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
	Ours	43.3	49.7	**44.2**	49.1	**51.3**	**62.6**	**46.0**	**44.9**	**58.2**	**65.7**	**49.9**	**47.3**	**51.3**	37.7	**40.0**	**49.4**

Experiment results

We report quantitative and qualitative results on *Human 3.6 M* and *MPI-INP-3DHP* datasets with the respective input ground-truth and estimated 2D pose

Table 2. Quantitative comparisons on MPI-INF-3DHP. All models are trained by 2D ground truth. T denotes the number of frames. Bold indicates best-to-date, and underline indicates second-best for each table respectively.

Venue	Method	T	PCK ↑	AUC ↑	MPJPE ↓
CVPR'19	Pavllo et al. [9]	81	86.0	51.9	84.0
CVPR'22	Li et al. [19]	9	93.8	63.3	58.0
CVPR'22	Zhang et al. [10]	1	94.2	63.8	57.9
CVPR'22	Zhang et al. [10]	27	94.4	66.5	54.9
CVPR'23	Gong et al. [14]	81	**98.0**	75.9	29.1
	Ours	1	**98.0**	**76.2**	**29.0**



a) Human3.6M

b) MPI-INF-3DHP

Experiment results

- We report the inference speed on an Nvidia A100 GPU.
- Despite the proposed Cross-Channel Embedding (CCE) module and the Context Attention Guidance (CAG) module introducing increased computing complexity, compared to other approaches, we have achieved an improvement in performance. Correspondingly, the inference speed of our model has experienced only a negligible decline.

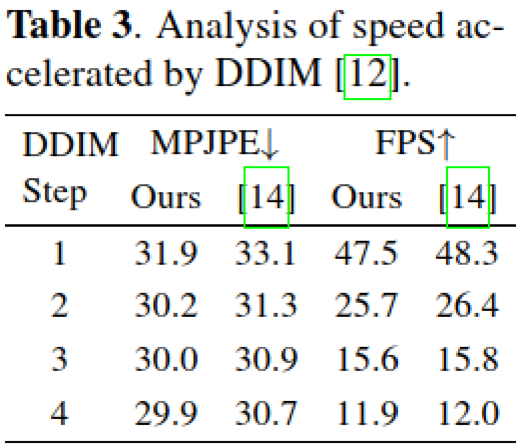


Table 3. Analysis of speed accelerated by DDIM [12].

DDIM	MPJPE↓		FPS↑	
Step	Ours	[14]	Ours	[14]
1	31.9	33.1	47.5	48.3
2	30.2	31.3	25.7	26.4
3	30.0	30.9	15.6	15.8
4	29.9	30.7	11.9	12.0

Summary

Conclusion

- A monocular 3D pose prediction framework.
- A novel cross-channel embedding module, to explore the correlation between joint-level 2D and 3D features.
- A context guidance mechanism, to facilitate the propagation of joint graph attention.
- Comprehensive evaluations demonstrated a significant improvement in terms of reconstruction accuracy compared to state-of-the-art methods.

To Learn more about our paper

- Project page: https://jjkislele.github.io/pages/projects/monoMotionDiff/
- Contact me: csjkjiang@comp.hkbu.edu

Exploring Latent Cross-Channel Embedding for Accurate 3D Human Pose Reconstruction in a Diffusion Framework

Q&A

Speaker: JIANG, Junkun
csjkjiang@comp.hkbu.edu.hk

2024/04/18