MuPPet: Multi-person 2D-to-3D Pose Lifting

Thomas C. Markhorst[1], Zhi-Yi Lin[1], Jouh Yeong Chew[2], Jan C. van Gemert[1], Xucong Zhang[1]
[1]Delft University of Technology, [2]Honda Research Institute Japan

1. Problem statement

- Main problem: estimating 3D pose of multiple humans in a social interaction (hard due to occlusions)
- Objective: leverage well performing 2D pose detections & inter-person correlations to lift 2D-to-3D pose

2. Contributions

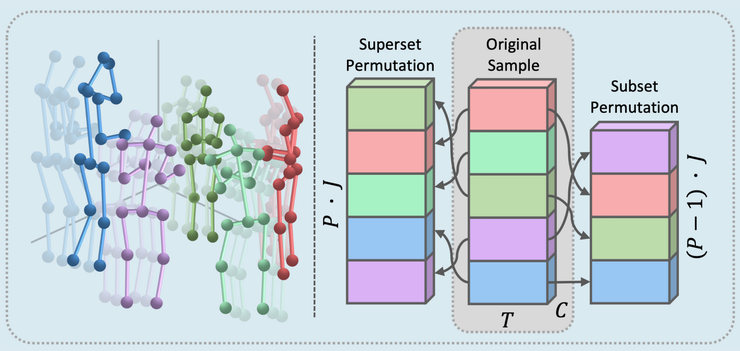
1. Novel method for pose lifting a dynamic number of persons in a group interaction across frames to 3D
2. Efficient intra- and inter-person modeling with proposed Person Encoding & Permutation Learning
3. Improved performance and occlusion handling compared to single-person baselines

3. Method

1. Architecture:

- Diffusion pipeline with transformer backbone (Fig. 3)
  1. Conditioning using 2D pose X
     - Denoise Gaussian noise 3D pose sample Y_T
  2. Temporal & Multi-Person Spatial Attention
     - Temporal modeling for consistency and counter occlusion
     - Multi-Person Spatial modeling leverages correlations between the joints of engaged persons
     - Person Encoding ensures distinguishability of different persons in attention modules

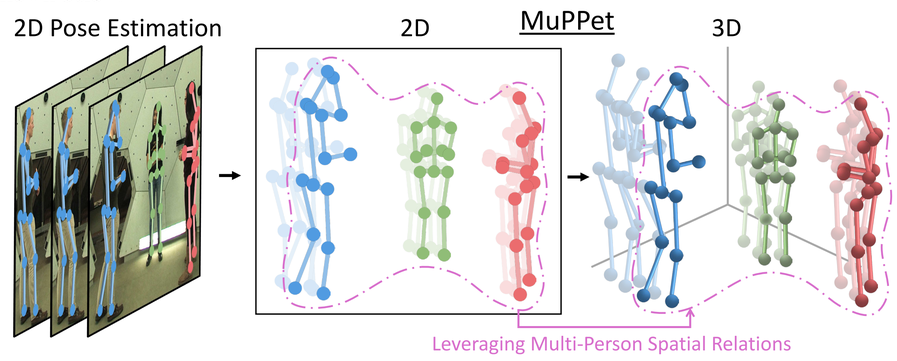
Fig. 2: Example of permutation augmentation for a five-person scene. We illustrate the people with color coding on the left and the permutation in the feature space on the right. From the Original Sample in the middle, we show a possible superset permutation on its left, and a possible subset permutation on its right.



2. Training

- Permutation Learning improves data variety while respecting inter-person relations (Fig. 2)
  1. Permute the order of persons in interaction
  2. Subsample & permute the order of persons in interaction to capture interactions in sub-groups
- Training loss: balance absolute & relative pose loss

Fig. 1: We exploit social inter-person correlations for 3D pose estimation to help infer unseen, occluded poses and spatial relations between individuals. Our method takes a sequence of 2D body poses to predict the sequence of 3D poses as shown below.



4. Results

CMU Panoptic Dataset [52]:

1. MuPPet handles multi-person scenario better than single-person method D3DP

Tab. 1: Comparison of MuPPet with single-person lifting method D3DP on the Haggling dataset, in absolute $\text{MPJPE}_{\text{abs}}$, relative $\text{MPJPE}_{\text{rel}}$, and absolute root $\text{MPJPE}_{\text{root}}$ pose estimation in mm. Our MuPPet achieves better performance than the SOTA single-person pose lifting method D3DP.

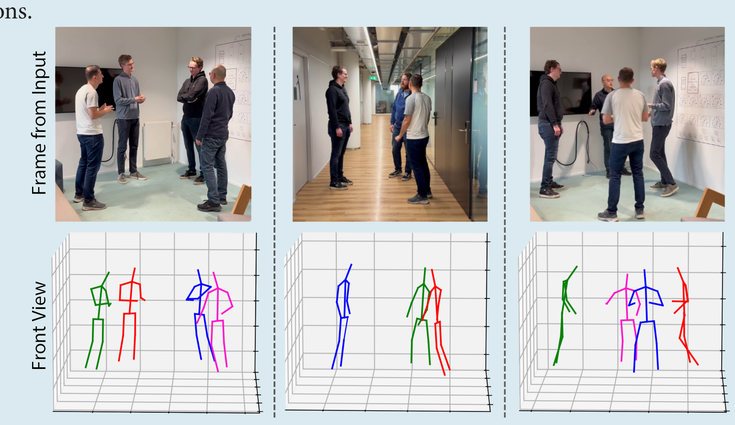
Method	$\text{MPJPE}_{\text{rel}}$ ↓	$\text{MPJPE}_{\text{abs}}$ ↓	$\text{MPJPE}_{\text{root}}$ ↓
D3DP [32]	58.2	-	-
$\text{D3DP}_{\text{absolute}}$ [32]	59.1	144.4	135.9
MuPPet	**55.3**	**119.5**	**108.9**

2. Better mean performance and lower variance than multi-person competitors

Tab. 2: Comparing our method with SOTA Multi-Person pose lifting (VirtualPose and POTR-3D) and direct estimation methods (MubyNet and SMAP) on the CMU Panoptic dataset. All numbers are the averaged joint error in mm. We only list the relative joint error $\text{MPJPE}_{\text{rel}}$ since the absolute joint error $\text{MPJPE}_{\text{abs}}$ is not reported by previous pose lifting methods.

Method	Haggling	Mafia	Ultimatum	**Mean**
MubyNet [56]	72.4	78.8	66.8	72.7
SMAP [22]	63.1	60.3	56.6	60.0
VirtualPose [14]	**54.1**	61.6	**54.6**	56.8
POTR-3D [19]	60.0	57.0	55.5	57.5
MuPPet	56.1	**54.3**	57.1	**55.8**

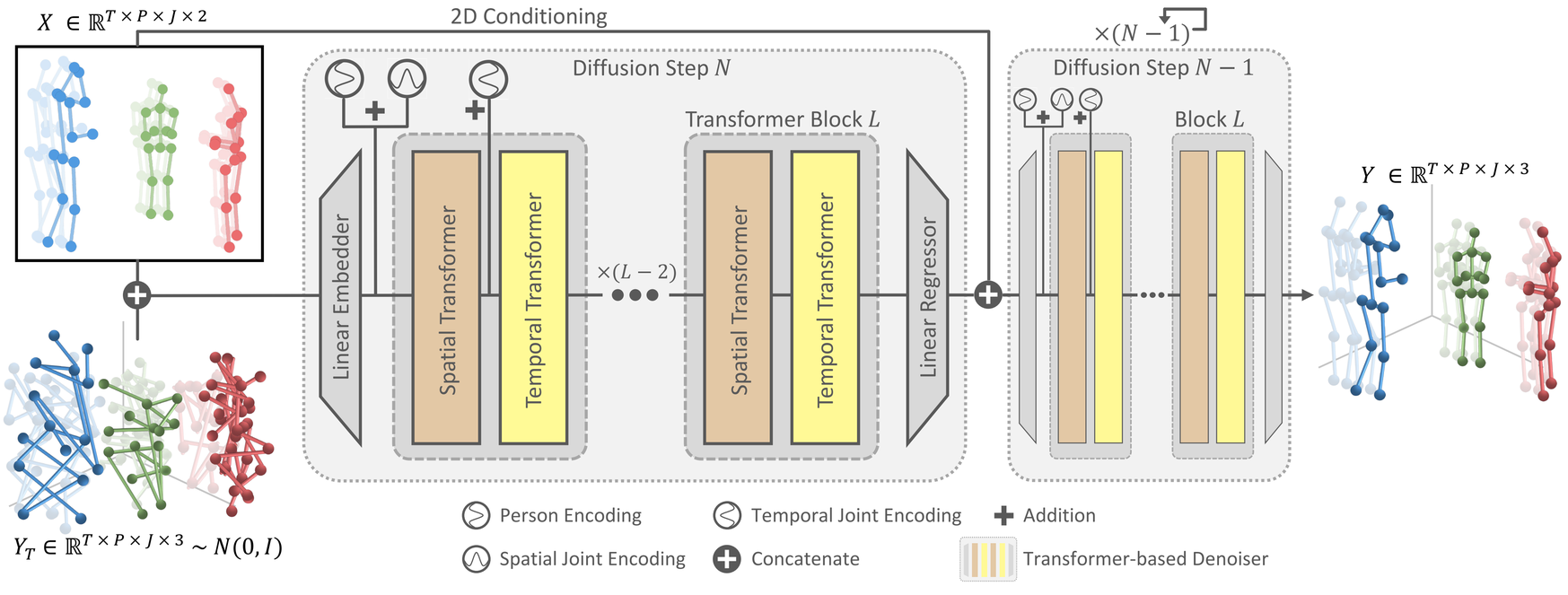
Fig. 4: Qualitative results on an in-the-wild setting predicted by MuPPet. We show one frame from the video input and the corresponding three views of the predicted 3D pose from multiple persons in the scene. Our MuPPet demonstrates effective performance in absolute 3D joint prediction, even on highly occluded persons.



5. Limitations & Conclusion

- Limitations: optimized specifically for group interaction
- Propose a novel 2D-to-3D multi-person pose-lifting method

Fig. 3: Overview of our MuPPet pipeline. Given a sequence of detected 2D human pose joints from multiple persons X, we use the diffusion process N times to denoise the 3D random poses Y_N to the output absolute 3D pose Y. Inside the denoiser, the spatial transformer and person encoding are applied to capture intra- and inter-person relationships, and a temporal transformer is used to capture the joint relationship across frames.



[32] W. Shan, Z. Liu, X. Zhang, Z. Wang, K. Han, S. Wang, S. Ma, and W. Gao, "Diffusion-based 3d human pose estimation with multihypothesis aggregation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14 761–14 771.
[56] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu, "Deep network for the integrated 3d sensing of multiple people in natural images," in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/6a6610feab86a1f294dbbf5855c74af9-Paper.pdf
[22] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, and X. Zhou, "SMAP: Single-shot multi-person absolute 3d pose estimation." [Online]. Available: http://arxiv.org/abs/2008.11469
[14] J. Su, C. Wang, X. Ma, W. Zeng, and Y. Wang, "Virtualpose: Learning generalizable 3d human pose models from virtual data," in European Conference on Computer Vision. Springer, 2022, pp. 55–71.
[19] S. Park, E. You, I. Lee, and J. Lee, "Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14 772–14 782.
[52] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction." [Online]. Available: http://arxiv.org/abs/1906.04158