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

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

Abstract

Multi-person social interactions are inherently built on coherence and relationships among all individuals within the group, making multi-person localization and body pose estimation essential to understanding these social dynamics. One promising approach is 2D-to-3D pose lifting which provides a 3D human pose consisting of rich spatial details by building on the significant advances in 2D pose estimation. However, the existing 2D-to-3D pose lifting methods often neglect inter-person relationships or cannot handle varying group sizes, limiting their effectiveness in multi-person settings. We propose MuPPet, a novel multi-person 2D-to-3D pose lifting framework that explicitly models inter-person correlations. To leverage these inter-person dependencies, our approach introduces Person Encoding to structure individual representations, Permutation Augmentation to enhance training diversity, and Dynamic Multi-Person Attention to adaptively model correlations between individuals. Extensive experiments on group interaction datasets demonstrate MuPPet significantly outperforms state-of-the-art single- and multi-person 2D-to-3D pose lifting methods, and improves robustness in occlusion scenarios. Our findings highlight the importance of modeling inter-person correlations, paving the way for accurate and socially-aware 3D pose estimation. Our code is available at: https://github.com/Thomas-Markhorst/MuPPet

1. Introduction

Nonverbal cues such as body pose are closely linked to group dynamics [13] and internal human states, including intention and emotion [69]. These cues also play a key role in effective human communication [16, 26, 33]. Therefore, accurately detecting the human pose of each person within a multi-person interaction is crucial for interpreting social cues. It has been shown that in a group activity, the head and body orientation [2, 61], individual body actions [4], and body pose associated with facial expression [39], are important for understanding social interaction and group dynamics. In addition, the perception of multi-person body motion is desirable for intelligent systems such as humanoid robots to interact with groups [1]. Noticeably, 3D human pose in absolute space is preferred over relative space in multi-person pose estimation since absolute pose can relate the positions and orientations between individuals in the real-world space, which is important for social interaction analysis [2, 61].

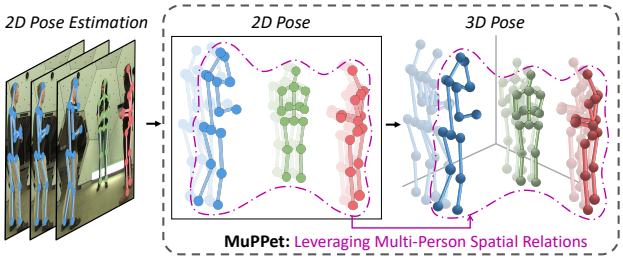

Figure 1. We exploit social inter-person correlations for 3D pose estimation to help infer occluded poses and spatial relations between individuals in a group interaction. Our method takes a sequence of detected 2D body poses to predict the sequence of 3D poses, as shown in the figure.

Despite the blooming development of 3D human pose estimation [36, 46, 60, 67], most existing works neglect the correlations between persons in social interaction. It is noteworthy that the correlations of multi-person body movements during group social interactions have been widely discussed as "imitation", "mimicry", "synchrony" [54]. It is promising to leverage such interactions between individuals in a group to enhance pose estimation performance in multi-person scenes. Unfortunately, most 3D pose estimation methods focus on single-person scenarios [41, 63, 75]. Although there are previous studies on multi-person 3D pose estimation [55–57], they mainly consider the crowded setting in terms of occlusion or depth placement and ignore any interaction between people. POTR-3D [43] is an exception as they model inter-person interaction. However, this method is trained with fixed group size [43], which is impractical in real-world scenarios.

In this paper, we propose a 2D-to-3D body pose lifting method MuPPet specifically designed for the multi-person

setting, which models the correlation information between persons via spatial attention, person encoding, and permutation learning. We pick the 2D-to-3D pose lifting approach due to its promises of leveraging accurate 2D pose estimation and focusing on the significance of the spatial 3D pose details [34, 43, 47, 68]. An overview of MuPPet is shown in Fig. 1, which takes the detected sequence of 2D poses as input and lifts them to the absolute 3D body pose in the world coordinate system.

Extending from the typical self-attention module in a single person [68], we apply the self-attention on the collection of persons to capture the relationship between people. We incorporate a person encoding into our model to associate each body joint with their corresponding individual explicitly, which causes the model to efficiently learn inter-person relationships. Building on this person encoding, we introduce a novel data augmentation strategy that permutes person ordering during training to enhance training diversity. The combination of spatial attention and person encoding enables MuPPet to handle arbitrary numbers of people in the scene, instead of a fixed maximum number of people [43]. We implement our approach using a diffusion process with a transformer backbone and include the temporal information across frames in a video.

Through extensive experiments, we demonstrate that the proposed method not only outperforms the single-person baseline but also exceeds the performance of current state-of-the-art 2D-to-3D multi-person lifting methods. Notably, our method exhibits clear advantages in handling occlusions compared to the single-person method. We believe this approach represents an important utility for advancing human behavior analysis in multi-person settings. In summary, our contributions are three-fold:

- A method for pose lifting a dynamic number of persons in a group interaction across frames to 3D.
- Efficient intra- and inter-person modeling with proposed person encoding, permutation learning, and spatial attention across all persons.
- Better performance and occlusion handling compared with a single-person baseline, and outperforms state-of-the-art (SOTA).

2. Related work

3D Pose Estimation. Estimation of a 3D human pose from a 2D image is a long-standing challenging problem [49, 63]. A straightforward solution for this task is directly taking the input to predict 3D pose [8, 11, 15, 38, 40, 73]. Prior knowledge of human kinematics has been investigated in previous works [45]. Especially, the SMPL body model [31] is commonly used in these 3D pose estimations [52, 56, 65, 70]. While these approaches can handle a variety of scenes, they require accurate 3D annotations, which is expensive and labor-intensive. In contrast, other methods [34, 43, 47, 68] lift 2D pose detections produced by 2D pose estimators trained on large datasets to 3D pose. These lifting methods reduce the need for 3D annotations in diverse environments and have recently gained popularity. Various architectures have been explored for single-person lifting, including transformers [28, 29, 35, 36, 47, 68], fully connected networks [12], graph convolutional networks [27, 77], and diffusion models [10, 22, 50, 64]. However, due to the limited 3D annotated data, data augmentation is commonly used and shows effectiveness for the pose lifting task [43, 46]. In this paper, we focus on 2D-to-3D lifting and propose a diffusion-based approach with a transformer-based denoiser.

A second distinction in 3D pose estimation is the use of temporal information [20]. Single-frame methods (frame2frame) [7] rely solely on per-frame inputs, making them sensitive to occlusions and noise when applied to a video. Sequence-to-frame methods [30, 44] improve robustness by predicting pose on one frame using a sequence of past and future frames. Moreover, sequence-to-sequence models [58, 68, 71, 74, 76] enforce temporal consistency across frames and enable more efficient inference. Based on the temporal handling advantages, we incorporate both input and output sequences in our approach.

Multi-person Pose. Multi-person 3D pose estimation introduces two primary challenges: placing poses in an absolute space and handling occlusions caused by other individuals. Direct 3D methods, such as BEV [56], excel at absolute positioning by explicitly modeling a bird's-eye-view representation. ROMP [57] incorporates a collision-aware strategy to mitigate inter-person overlap, while 3DCrowdNet [9] enhances robustness in crowded scenes by integrating an additional 2D pose extractor. However, these methods are frame-to-frame due to the high computational cost of end-to-end video processing. Moreover, they require extensive diverse training data to cover variances in appearance. To reduce dependency on diverse 3D training data, 2D-to-3D lifting approaches have also been explored. VirtualPose [55] predicts 3D pose using detected bounding boxes and a joint heat map, removing the image input from the 3D estimation. In contrast, POTR-3D [43] handles occlusion better due to its sequence-to-sequence modeling and leveraging of ground-truth data during training. Unfortunately, it is trained with fixed numbers of people in a scene and pads less crowded frames with zeros. Additionally, POTR-3D [43] applies data augmentations that separately translate and rotate individual persons. Such data augmentations bring performance improvements, however, they disrupt the inter-person relationships. To form a multi-person augmentation leveraging inter-person relationship, we propose the person encoding and then permute the person ordering to simulate various multi-person settings.

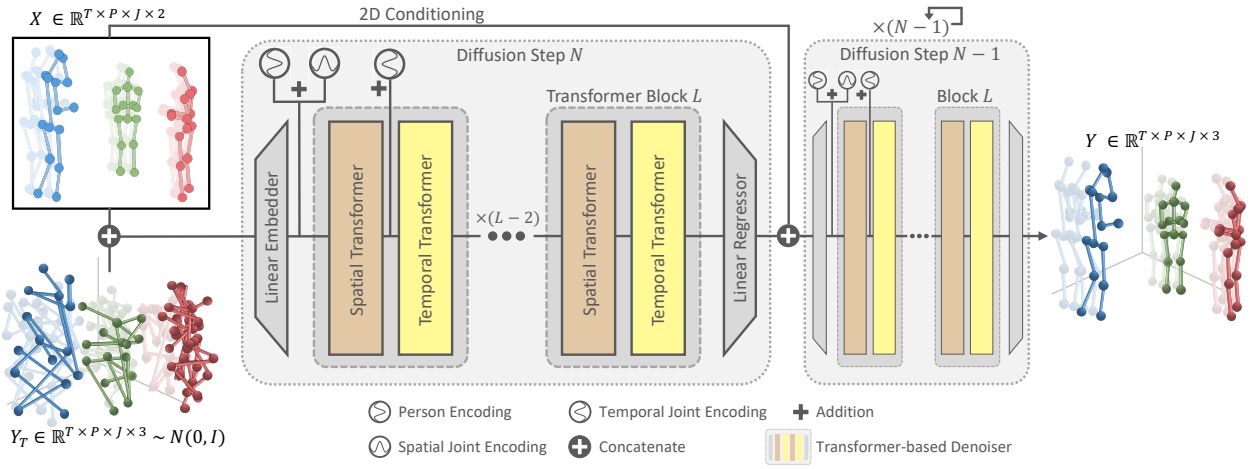

Figure 2. Overview of our MuPPet pipeline. Given a sequence of detected 2D human pose joints from multiple persons X, we use the diffusion process N times to denoise the 3D random poses Y_N to the output absolute 3D pose Y. Inside the denoiser, the spatial transformer and person encoding are applied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

Diffusion. Diffusion models, first introduced by [51] and specifically the Denoising Diffusion Probabilistic Models (DDPM) [17], have demonstrated strong performance across various generative tasks [3, 5, 18, 21, 42, 48]. Their probabilistic nature makes them well-suited for addressing key challenges in 3D human pose estimation such as occlusion and depth ambiguity. Given any 2D input, deterministic models must commit to a single prediction, potentially losing plausible alternatives, especially in occluded cases. In contrast, diffusion models can probabilistically reason over multiple valid 3D poses, generating diverse hypotheses instead of making a fixed choice. This has led to a few single-person pose-lifting methods using diffusion [50, 68]. These methods generate hypotheses, which can then be aggregated to refine final pose predictions and improve overall performance [19, 50], making diffusion-based approaches a promising direction for 3D pose estimation.

3. Method

Our target scenario is a multi-person social setting where the individuals are engaged in one activity, for example, a group dialogue. MuPPet takes a sequence of 2D body poses from multiple persons to output the lifted 3D human poses of all individuals in the sequence.

3.1. Architecture

The overview of the proposed method is shown in Fig. 2. Our method takes a detected sequence of 2D body poses $X \in \mathbb{R}^{T\times P\times J\times 2}$ of T frames, with P persons, J joints, and 2D joint positions as input. A linear layer maps X to an embedding $\hat{X} \in \mathbb{R}^{T\times P\times J\times 512}$, which is fed into an attention block with spatial and temporal attention modules. The attention block is repeated L times to generate the features that are fed into a regression head that maps the final attention output $\hat{X_L}$ to 3D multi-person pose $Y \in \mathbb{R}^{T\times P\times J\times 3}$. We employ the diffusion process similar to previous works [10, 22, 50, 64] to gradually generate the 3D pose with N times denoising process.

Multi-person spatial attention. Inspired by [68] that performs the attention within a single person, we extend it to capture all joints across any number of persons in the scene. Given the encoded feature from a frame t as $f_t \in \mathbb{R}^{P\times J\times 512}$, we reshape the representation to $\hat{f}_t \in \mathbb{R}^{(P\cdot J)\times 512}$, where P is a variable to handle varying group sizes. The query, key and value matrix Q, K, V are computed using collection of tokens $\hat{f}_t$ and weights W^K, W^Q, W^V. Followed by a linear projection W^L. Note all learned weights W^K, W^Q, W^V, W^L have dimension $\mathbb{R}^{512\times 512}$ which is independent of the amount of persons P and can therefore handle dynamic group sizes. In this way, we manage to apply self-attention across joints from all persons in the scene. To enable the model to handle varying numbers of persons, the model is trained with varying P, ensuring that both inter-person and intra-person spatial relations are captured for all group sizes.

Temporal attention. As our method is sequence-to-sequence, it should relate each joint across all frames. To achieve this, we first collect a sequence of tokens $s_{p,j} \in \mathbb{R}^{T\times 512}$. We then perform self-attention with $s_{p,j}$ with a temporal embedding similar to [68]. The query, key and value matrix Q, K, V are computed using collection of tokens $\hat{s}_{p,j}$ and weights W^K, W^Q, W^V. Followed by a lin-

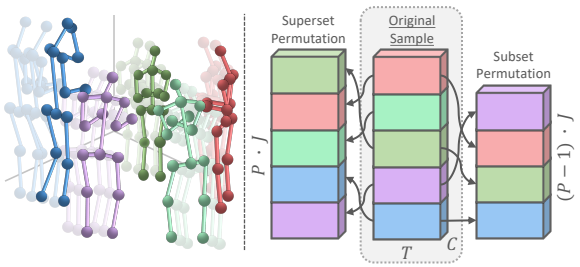

Figure 3. Example of permutation learning for a five person scene. We illustrate the people with color coding on the left and the permutation in the feature space on the right. From the original feature $\hat{X} \in \mathbb{R}^{T\times(P\cdot J)\times C}$, where C is the number of channels, we show a possible superset permutation of P persons with all J joints, and subset permutation of $(P-1)$ persons with all J joints. During training, n_{sub} and n_{sup} permutations are used.

ear projection W^L. Following the attention operation, the representations $s_{p,j}$ for all persons p and joints j are concatenated and reshaped to form $\hat{X}_l \in \mathbb{R}^{T\times P\times J\times 512}$, which serves as the output of the spatio-temporal block.

3.2. Person Encoding

Both the spatial and temporal transformer modules do not distinguish which person a joint belongs to. This introduces challenges to using the inter-person information for 3D pose lifting. To explicitly encode the person information into the model, we propose the person encoding $E \in \mathbb{R}^{P\times 512}$, which is the same for all joints $j \in J_p$ that belong to person p. E is a set of learned parameters. The person encoding ensures the model distinguishes the joints that belong to individuals while considering the relationship between people. It addresses the issue in previous multi-person pose lifting [43] that can only process the fixed maximum number of persons and add zero padding to missing persons. Together with the spatial attention across all joints, the person position embedding grants the capability of our method to handle multi-person sequences and, importantly, dynamic group sizes. Specifically, E is added to $\hat{X}$ before the first spatial transformer block as $\hat{X}_{p,j} = \hat{X}_{p,j} + E_p$ for all $p \in P$ and $j \in J$. Similar to the spatial and temporal joint encoding, E is added to $\hat{X}$ again at every diffusion timestep.

3.3. Permutation Learning

As the multi-person pose lifting requires the consideration of the relationship between people, the commonly used data augmentations in the single-person setting, such as translation and rotation, cannot be applied. Interestingly, our newly proposed person encoding allows us to perform a new permutation data augmentation specifically for the multi-person 3D pose lifting task.

Given an input-output pair X, Y which consists of P persons, we highlight that there is no specific ordering of the individual persons in P. Consequently, arbitrarily permuting the order of P fed into the model could increase the diversity of the data. However, the spatial and temporal attention modules themselves are invariant to such person order permutations since they do not distinguish individuals. Fortunately, our person encoding explicitly assigns the embedding E_p to each person p in the scene to enable the model to distinguish individuals between permutations. We exploit this by randomly permuting the order of P of any sample X, called *superset permutation*. The person encodings are added to samples only after permuting. We illustrate the permutation learning with an example in Fig. 3.

Additionally, we hypothesize that a subset of P in a social setting exhibits similar relations as its superset. Therefore, we also take varying subsets of P and randomly permute them, which we call *subset permutation*. Taking both strategies together, we result in the permutation learning approach where for each sample X we train on i) X itself in normal person order, ii) n_{sup} random permutations of the person order of X and iii) n_{sub} evenly divided over the two subset sizes $|P|-1$ and $|P|-2$. We define n_{sup} as the number of permutations of the superset, and n_{sub} as the number of permutations for the subset.

3.4. Diffusion Process

Following DDPM [17], we employ a diffusion process to gradually corrupt our multi-person 3D pose $Y \in \mathbb{R}^{T\times P\times J\times 3}$ by adding Gaussian noise over N time steps, such that Y_N becomes nearly pure Gaussian noise. The transformer architecture described in Section 3.1 is used as a denoiser in the reverse diffusion process. The process is defined in Eq. 1, where $\bar{\alpha}_t \in [0,1]$ is a fixed hyperparameter to control the noising scheme and $Y_0 = Y$:

$$Y_n = \sqrt{\bar{\alpha}_n} Y_0 + \sqrt{1-\bar{\alpha}_n}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

We condition the reverse diffusion process on detected 2D pose X. Following [50], we concatenate X with Y_n to form $Z_n \in \mathbb{R}^{T\times P\times J\times 5}$ at diffusion step n. From every Z_n the denoiser outputs $\tilde{Y}_0$ targeting Y_0, following DDIM [53]. With any $n' < n$, $Y_{n'}$ is constructed by adding noise as shown in Eq. 2. Then $Y_{n'}$ is again concatenated with X to construct $Z_{n'}$.

$$Y_{n'} = \sqrt{\bar{\alpha}_{n'}} \cdot \tilde{Y}_0 + \sqrt{1-\bar{\alpha}_{n'}} \cdot \epsilon + \sigma_n \epsilon \tag{2}$$

$$\tilde{Y}_0 = \text{denoise}(\text{concat}(X, Y_n)) \tag{3}$$

At inference, We sample multiple Y_N conditioning on the same X, which gives multiple outputs. We then aggregate these by taking the joint-wise average, to boost performance as in previous works [19, 50].

3.5. Pose Loss

We experimentally find that directly outputting absolute positions of all joints is a difficult task. Therefore, instead of predicting all joints of Y in absolute space, we normalize poses by separating relative pose and absolute root location. For the relative pose, each joint is transformed into a root-relative coordinate system, where the hip-center serves as the root joint, following standard practice [59, 68]. The root-relative positions typically fall within $[-1, 1]$ meters, requiring no further normalization. However, the absolute root joint is normalized based on the training set as it has a larger range. The single absolute root joint and remaining relative joints have separate loss terms $\mathcal{L}_{\text{abs}}$ and $\mathcal{L}_{\text{rel}}$, respectively. To balance the weight between relative and absolute joint calculation, we introduce a weight λ:

$$\mathcal{L}_{\text{MPJPE}} = \lambda \cdot \mathcal{L}_{\text{abs}} + (1 - \lambda) \cdot \mathcal{L}_{\text{rel}} \tag{4}$$

4. Experiments

We first compare our multi-person method MuPPet with the SOTA single-person 3D pose lifting method to highlight its advantages in the multi-person setting. We then compare MuPPet to the current SOTA multi-person 3D pose lifting methods with the reported numbers in their original papers. We further show the effectiveness of MuPPet in handling occlusions. Finally, we conduct ablation studies to analyze the impact of individual modules in our proposed method.

4.1. Experimental Settings

Datasets The Haggling dataset [23] contains 30 recordings with 173 separate sequences inside a capture system with 31 cameras. Within each recording, three participants play a haggling activity for around one minute. Each person is annotated with 19 body pose joints following the COCO19 format. It captures social interaction [59] due to participants having specific social roles compared to other datasets [37, 62], which increases its relevance for our study. Models evaluated on the Haggling dataset are trained on 133 sequences, the corresponding test set contains 40 sequences. We use six superset and six subset permutations due to having only three persons in each scene.

CMU Panoptic [24] is an older dataset than Haggling [23] and is recorded in the same capture system. It contains several social games varying in group size from two to eight persons with 13 recordings and 58 separate sequences, significantly less compared to Haggling [23]. Different from the 19 body joints annotated in the Haggling dataset, Panoptic has 15 body joints following the MPI15 format. Following [43, 55, 73], we train and evaluate on cameras with the index of 16 and 30. We focus on the Haggling, Mafia, and Ultimatum scenes in the dataset, as these sessions contain a single group interaction. Other scenes, like Pizza, contain either fragmented groups or single persons and are therefore not suitable for our group based lifting method. The model evaluated on Panoptic is trained on 38 sequences and tested on 20 sequences. Due to the low number of training samples, models tested on Panoptic are pre-trained on Haggling and finetuned with half the initial learning rate. Due to the small size of this dataset and having more subjects compared to the Haggling dataset, we set the permutation to be 13 superset and three subset permutations.

Implementation Details. We set $L = 8$ such that the network contains eight spatio-temporal blocks. The network is optimized for 400 epochs using AdamW [32] in PyTorch, with a starting learning rate of 6E-5 and exponential decay of 0.997 per epoch. The maximum timesteps for the diffusion process are set to 1000, the batch size is four, and we train on pose sequences of 243 frames following [68]. Training and inference are run on a single NVIDIA A100 GPU. For all models, the input 2D pose sequences are obtained by OpenPose detection [6], and person-id is matched with ground-truth.

Metrics. Following [59] for the Haggling dataset, and [43, 55, 56] for the CMU Panoptic dataset, we report mean per joint positional error (MPJPE) in relative space and absolute space with mm as a unit. We only consider joints that are inside the camera frame. For the relative joint error calculation, we align the predicted and GT root joint and average the error. For the absolute joint error calculation, we take the global origin and calculate the average joint error relative to the origin. The root error is the error in absolute space of only the root joint estimation.

4.2. Comparison of Multi-person and Single-person

To reveal the benefit of correlation learning between persons, we conduct experiments on the Haggling dataset due to the rich interactions between people. We pick the SOTA single-person 3D pose lifting method D3DP [50] as the baseline in this experiment, which is similar to our method in terms of using diffusion for human pose lifting, while only processing a single person individually. We only compare the D3DP since it shows superior performances compared with current other single-person 3D pose lifting methods in the original paper [50]. The original D3DP is trained only on the relative joints, we alter it to predict both relative and absolute joint positions ($\text{D3DP}_{\text{absolute}}$). Note that we evaluate D3DP with the exact same training and test sets as MuPPet for a fair comparison.

In Tab. 1, we can see that MuPPet outperforms the D3DP baseline in both relative and absolute joint position errors, with significant margins of 6.4% (from 59.1 to 55.3) in the relative, 17.2% (from 135.9 to 108.9) in the absolute joint position errors, and 19.9% in the absolute root joint error. It clearly demonstrates the benefits of leveraging intra- and inter-person relationships for the 3D pose-lifting

Method	$\text{MPJPE}_{\text{rel}} \downarrow$	$\text{MPJPE}_{\text{abs}} \downarrow$	$\text{MPJPE}_{\text{root}} \downarrow$
D3DP [50]	58.2	-	-
$\text{D3DP}_{\text{absolute}}$ [50]	59.1	144.4	135.9
MuPPet	**55.3**	**119.5**	**108.9**

Table 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.

task. The improvement in absolute joint error is expected as we hypothesized that modeling multi-person relations gives an improved understanding of 3D location. We attribute the improvement in relative joint error to our permutation learning approach used as data augmentations, which cannot be applied to single-person methods. Note that the performance improvement of the absolute joint error is higher than the relative (17.2% vs 6.4%), which strongly indicates the benefit of MuPPet in handling the absolute joint estimation. This is relevant, as the absolute joint position is more meaningful than the relative joint position in the multi-person interaction to understand the group dynamics.

By comparing the original D3DP and our alternation with absolute joint loss, we can see that adding the absolute joint loss slightly decreases the relative pose performance. Which could be caused by predicting relative and absolute joint positions jointly being a difficult task.

4.3. Comparison with SOTA

It is difficult to compare with the previous SOTA methods due to the limited datasets with group interaction, availability of source code from previous works, and the lack of absolute joint error metric. In this section, we make an effort to compare MuPPet with the current SOTA 3D pose estimation method including 2D-to-3D pose lifting and direct 3D pose estimation. We use Panoptic in this experiment as it is popular for the 3D pose lifting task. We report only the relative joint error due to missing absolute joint error results. We focus on the Haggling, Mafia, and Ultimatum scenes in Panoptic, as they fit our target setting where all people engage in one group activity. Since MuPPet is a 2D-to-3D multi-person pose lifting method, we mainly compare it against VirtualPose [55] and POTR-3D [43]. We also list the performances reported by other SOTA multi-person direct 3D pose estimation models [66, 73].

The results are shown in Tab. 2, with performances in individual scenes and averaged errors across Haggling, Mafia, and Ultimatum scenes. We can see from the table that multi-person 2D-to-3D pose lifting methods, *i.e*. VirtualPose [55], POTR-3D [43], and our MuPPet, achieve better performances than direct estimation methods. Among

Method	Haggling	Mafia	Ultimatum	**Mean**
MubyNet [66]	72.4	78.8	66.8	72.7
SMAP [73]	63.1	60.3	56.6	60.0
VirtualPose [55]	**54.1**	61.6	**54.6**	56.8
POTR-3D [43]	60.0	57.0	55.5	57.5
MuPPet	56.1	**54.3**	57.1	**55.8**

Table 2. Comparing our method with SOTA Multi-Person pose lifting (VirtualPose and POTR-3D) and direct estimation methods (MubyNet and SMAP) on the on the CMU Panoptic dataset. All numbers are 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. Our MuPPet achieves the best performance in the multi-person pose estimation setting.

the top three methods, our MuPPet achieves the best performance in terms of the averaged relative joint errors across the three scenes. Note that, although the optimization target of our method is the absolute joint pose, MuPPet still outperforms previous SOTA 2D-to-3D methods in the relative joint error. It shows the benefit of leveraging the intra- and inter-person relationships for multi-person pose estimation.

4.4. Occlusion study

In social interactions with multi-people, individuals often occlude each other, making pose estimation challenging. Unlike the single-person pose estimation method, our multi-person pose estimation approach can infer the occluded body pose from surrounding other people. We assume that social interactions exhibit coherence between individuals, and our method exploits it to improve pose estimation under occlusions. To validate this, we compare our method with the single-person baseline D3DP on the Haggling dataset, which has rich interaction among people. We summarize the results in the Fig. 4 with the relative joint error and absolute joint error computed over different numbers n of occluded joints per person, where n is the maximum number of occluded joints for one person.

The figure clearly shows that our method handles occlusion better than D3DP across any number of occluded joints. The performance gap becomes larger as the number of occluded joints increases, with the effect being particularly pronounced in the absolute joint error. Since we focus on improving the absolute joint estimation in multi-person settings, which are often occluded, the result indicates the effectiveness of MuPPet.

4.5. Ablation study

Component Ablation In this section, we examine the different components of the proposed method and their effect, using the Haggling dataset. We conduct ablation studies on dynamic multi-person handling, person encoding, permuta-

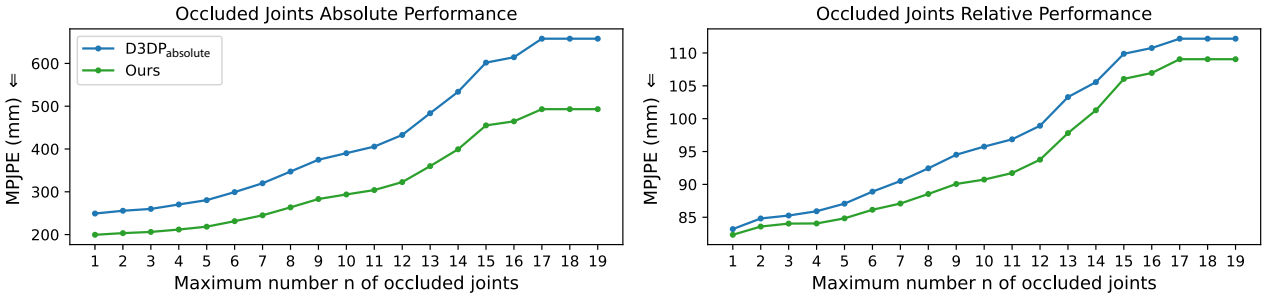

Figure 4. We show the performance of different levels of occlusion on the Haggling dataset, in comparison with the adapted D3DP singleperson model. The X-axis is the maximum number of occluded joints, and the Y-axis is the joint error in mm. Our method performs better in relative joint error MPJPE$_{\text{rel}}$ and signficantly better in absolute joint error MPJPE$_{\text{abs}}$ over all levels of occlusion.

Multi	PE	Sup.Perm.	Sub.Perm.	MPJPE$_{\text{rel}}$ ↓	MPJPE$_{\text{abs}}$ ↓
-	-	-	-	59.1	144.4
✓	-	-	-	68.4	142.1
✓	✓	-	-	69.1	131.5
✓	✓	✓	-	60.0	126.9
✓	✓	✓	✓	**55.3**	**119.5**

Table 3. Ablation showing that each module contributes to the performance. Dynamic multi-person attention (Multi), Person Encoding (PE), super set permutation (Sup.Perm, and subset permutation learning (Sub.Perm). We show both relative joint error MPJPE$_{rel}$ and absolute joint error MPJPE$_{abs}$.

# Supersets	# Subsets	MPJPE$_{\text{rel}}$ ↓	MPJPE$_{\text{abs}}$ ↓
12	0	58.4	151.5
6	6	56.5	150.0
2	10	**56.4**	**143.0**
1	7	56.3	148.6
2	10	56.4	**143.0**
3	13	**55.8**	146.8

Table 4. Ablation study on the permutation numbers in superset n_{sup} and subset n_{sub}. Upper: we fix the total number of permutations to be 12 and alter the distribution of n_{sup} and n_{sub}. Lower: we change the total number of permutations while keep the rough ratio of n_{sup} and n_{sub}. All numbers are joint errors in mm. We observe that adding subset permutations and increasing total permutations increases performance significantly.

tion learning, and subset permutation learning. In Tab. 3, we show the performances by gradually adding each component to our model. The first row in the table is the single-person baseline. By integrating our multi-person architecture, the absolute joint performance increased (from 144.4 to 142.1), however, at the cost of relative joint performance (from 59.1 to 68.4). Adding person encoding significantly improves the absolute joint performance to 131.5 while the relative joint performance gets slightly worse (from 68.4 to 69.1). Combined with superset permutation learning, absolute joint error decreases to 126.9 and relative joint estimation performance gets on par with the single-person baseline (60.0 vs 59.1). Finally, adding subset permutation learning decreases relative joint error to 55.3, and also decreases the absolute joint error such that our final model outperforms its other version by more than 5.8% (from 126.9 to 119.5).

Permutation ablation We propose two permutations, the superset and subset permutations, to improve the diversity of the training samples. To determine the effect of permutation learning and find the optimal number of permutations during training, we perform a study on different settings of permutations. We perform this ablation on the Panoptic dataset since it contains a various number of subjects to test the subset permutations. We report the joint error averaged over all four scenes. First, we test the balance between superset and subset permutations. We keep the total number of permutations as 12 due to the trade-off between performance and computation, and change the ratio of superset and subset permutations. We show the results in the upper part of Tab. 4, which indicate that using subset permutations has a significant benefit over only using superset permutations, as the difference between 12-0 and 6-6 ratios is most significant for the relative performance. We attribute this performance boost to the increase in the diversity of the number of people in each scene. The 2-10 ratio, shows a further improvement mainly in absolute pose.

We then examine the total amount of permutations and show results in the lower part of Tab. 4, which shows that increasing the number of permutations improves performance, as expected. However, the change becomes marginal or even negative with a high number of permutations, as the absolute joint error slightly increases when moving from the 2-10 ratio to the 3-13 ratio, despite a decrease in relative joint error.

Diffusion Process Following [50] we use the DDIM diffusion process, allowing us to run inference with only five steps rather than the 1000 steps the model is trained with. As there are arguments from previous works on whether the diffusion process is necessary [72], we evaluate the influence of our diffusion approach on the Haggling dataset. Therefore, we compare MuPPet and a version trained and tested without the diffusion process, *i.e*. only the transformer backbone itself. Our results show that integrating the multi-step diffusion process can improve the relative joint error from 75.0 to 55.3 mm and the absolute joint error from 167.5 to 119.5 mm. This shows the necessity of the diffusion process in our model.

4.6. Qualitative Results on Social Dataset

To showcase the performance of our method on in-the-wild social interaction, we qualitatively analyze MuPPet on in-the-wild social interaction. The videos are recorded with a camera from a cellphone in different scenes. We use OpenPose to perform the 2D pose detection on the images, and match persons across frames using the Euclidean distance and Hungarian matching [25]. MuPPet lifts these 2D detections to the absolute 3D joint positions. As shown in Fig. 5, MuPPet can successfully locate and place all people from the scene in the 3D world coordinate system, which can be used for social analysis such as F-formations [14], physical distance between people, and body orientation [2, 61]. Moreover, we observe that even significantly occluded persons are reconstructed well, as can be seen in the second and third columns. The video versions of these results are available in the supplementary material.

5. Discussion and Conclusion

Limitations Our method is optimized for a specific setting of multi-person interaction and absolute joint estimation. In addition, the people in the group interaction must be engaged in a joint activity to form the inter-person relationship that our model can leverage. Lastly, we should have the absolute 3D joint position of each person for training and evaluation. Given these requirements, there is limited data that we could train and evaluate our model on. Specifically, the MuPoTS-3D [37] dataset that is popularly used in previous works is not suitable for our model. Mainly due to the training data of MuPoTS-3D, which is artificially combined of multi-people from different scenes without any real interaction between them. Additionally, most evaluation scenes of MuPoTS-3D do not contain natural social interaction.

As for future work, we see an opportunity to explore the usage of audio information in group activities for the pose estimation task. Given the fast development of multi-modality learning, we believe it is feasible to add the audio modality to aid the pose inference, given the correlation between speech and body gesture.

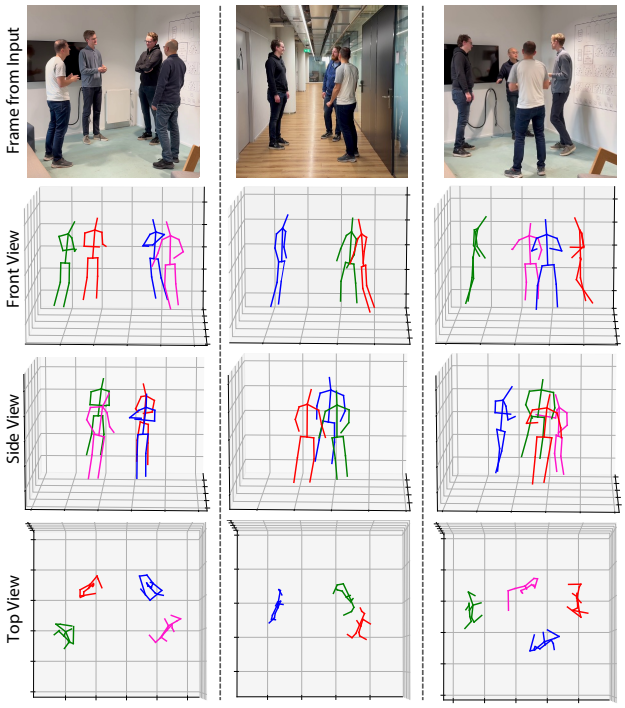

Figure 5. Qualitative results on an in-the-wild setting predicted by MuPPet. We show one frame from the video input and corresponding three views of the predicted 3D pose from multiple persons in the scene. Our MuPPet demonstrates effective performance in absolute 3D joint prediction, even on highly occluded persons.

Conclusion In this paper, we propose a novel 2D-to-3D human pose-lifting method MuPPet specifically for the multi-person group interaction setting. To learn the intra- and inter-person relationship, we propose dynamic spatial-temporal attention across all people, person encoding to distinguish each person, and permutation learning to increase the training data diversity. We successfully show the better performance of MuPPet compared to the SOTA pose lifting method, and demonstrate its advantages in handling occlusion scenarios. We further showcase the qualitative results of MuPPet on in the wild group social interaction data.

References

[1] Towards a humanoid museum guide robot that interacts with multiple persons. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pages 418–423. IEEE, 2005. 1

[2] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 5–14, 2015. 1, 8

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *CoRR*, abs/2111.14818, 2021. 3

[4] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and Francois Bremond. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 70–79, 2022. 1

[5] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *CoRR*, abs/2111.13606, 2021. 3

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[7] Ju Yong Chang, Gyeongsik Moon, and Kyoung Mu Lee. Absposelifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. *CoRR*, abs/1910.12029, 2019. 2

[8] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 2

[9] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. 3dcrowdnet: 2d human pose-guided3d crowd human pose and shape estimation in the wild. *CoRR*, abs/2104.07300, 2021. 2

[10] Jeongjun Choi, Dongseok Shim, and H. Jin Kim. DiffuPose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3773–3780. ISSN: 2153-0866. 2, 3

[11] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2271, 2019. 2

[12] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields, 2022. 2

[13] Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. Seeing is believing: body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)*, 29 (4):1–9, 2010. 1

[14] Susan Fiksdal. Conducting interaction: Patterns of behavior in focused encounters. adam kendon. cambridge: Cambridge university press, 1990. pp. vii+ 292. 16.95 paper. *Studies in Second Language Acquisition*, 15(1):116–117, 1993. 8

[15] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8575–8584, 2021. 2

[16] Judith A Hall, Terrence G Horgan, and Nora A Murphy. Nonverbal communication. *Annual review of psychology*, 70 (2019):271–294, 2019. 1

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. 3, 4

[18] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *CoRR*, abs/2106.15282, 2021. 3

[19] Karl Holmquist and Bastian Wandt. DiffPose: Multi-hypothesis human pose estimation using diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15931–15941. IEEE. 3, 4

[20] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 68–84, 2018. 2

[21] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech, 2022. 3

[22] Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6130–6140. IEEE. 2, 3

[23] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. 5

[24] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart C. Nabbe, Iain A. Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *CoRR*, abs/1612.03153, 2016. 5

[25] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. 8

[26] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 763–772, 2019. 1

[27] Han Li, Bowen Shi, Wenrui Dai, Yabo Chen, Botao Wang, Yu Sun, Min Guo, Chenglin Li, Junni Zou, and Hongkai Xiong. Hierarchical graph networks for 3d human pose estimation. *CoRR*, abs/2111.11927, 2021. 2

[28] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. MHFormer: Multi-hypothesis transformer for 3d human pose estimation. pages 13147–13156. 2

[29] Wenhao Li, Mengyuan Liu, Hong Liu, Pichao Wang, Jialun Cai, and Nicu Sebe. Hourglass tokenizer for efficient transformer-based 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 604–613, 2024. 2

[30] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned

multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[33] Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen, and Daniel Gatica-Perez. Body communicative cue extraction for conversational analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013. 1

[34] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2

[35] Soroush Mehraban, Vida Adeli, and Babak Taati. MotionAGFormer: Enhancing 3d human pose estimation with a transformer-GCNFormer network. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6905–6915. IEEE. 2

[36] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6920–6930, 2024. 1, 2

[37] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular RGB. 5, 8

[38] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017. 2

[39] Dimitris Metaxas and Shaoting Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31(6-7):421–433, 2013. 1

[40] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single RGB image. 2

[41] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *Ieee Access*, 8:133330–133348, 2020. 1

[42] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021. 3

[43] Sungchan Park, Eunyi You, Inhoe Lee, and Joonseok Lee. Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14726–14736. IEEE. 1, 2, 4, 5, 6

[44] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training, 2019. 2

[45] Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1123–1132, 2024. 2

[46] Qucheng Peng, Ce Zheng, and Chen Chen. A dual-augmentor framework for domain generalization in 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2240–2249, 2024. 1, 2

[47] Xiaoye Qian, Youbao Tang, Ning Zhang, Mei Han, Jing Xiao, Ming-Chun Huang, and Ruei-Sung Lin. HSTFormer: Hierarchical spatial-temporal transformers for 3d human pose estimation. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 3

[49] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016. 2

[50] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14715–14725. IEEE. 2, 3, 4, 5, 6, 8

[51] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. 3

[52] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 2

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. 4

[54] Darja Stoeva, Andreas Kriegler, and Margrit Gelautz. Body movement mirroring and synchrony in human–robot interaction. *J. Hum.-Robot Interact.*, 13(4), 2024. 1

[55] Jiajun Su, Chunyu Wang, Xiaoxuan Ma, Wenjun Zeng, and Yizhou Wang. VirtualPose: Learning generalizable 3d human pose models from virtual data. 1, 2, 5, 6

[56] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13233–13242. IEEE. 2, 5

[57] Yu Sun, Qian Bao, Wu Liu, Yili Fu, and Tao Mei. Centerhmr: a bottom-up single-shot method for multi-person 3d mesh recovery from a single image. *CoRR*, abs/2008.12272, 2020. 1, 2

[58] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-

temporal criss-cross attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4790–4799. IEEE. 2

[59] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9567–9577. IEEE. 5

[60] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[61] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci. Joint estimation of human pose and conversational groups from social scenes. *International Journal of Computer Vision*, 126(2):410–429, 2018. 1, 8

[62] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[63] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 1, 2

[64] Jinglin Xu, Yijie Guo, and Yuxin Peng. Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 561–570, 2024. 2, 3

[65] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7159–7169, 2021. 2

[66] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 6

[67] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3517–3526, 2019. 1

[68] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. MixSTE: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. 2, 3, 5

[69] Mingming Zhang, Yanan Zhou, Xinye Xu, Zhiwei Ren, Yihan Zhang, Shenglan Liu, and Wenbo Luo. Multi-view emotional expressions dataset using 2d pose estimation. *Scientific Data*, 10:649, 2023. 1

[70] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024. 2

[71] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8877–8886, 2023. 2

[72] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 8

[73] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3d pose estimation. 2, 5, 6

[74] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665, 2021. 2

[75] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 1

[76] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. version: 5. 2

[77] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11477–11487, 2021. 2