Capturing Closely Interacted Two-Person Motions with Reaction Priors

Qi Fang[1] Yinghui Fan[1] Yanjun Li[1] Junting Dong[2] Dingwei Wu[1]
Weidong Zhang[1] Kang Chen[1]
[1]NetEase Games AI Lab [2]Shanghai AI Lab

Abstract

In this paper, we focus on capturing closely interacted two-person motions from monocular videos, an important yet understudied topic. Unlike less-interacted motions, closely interacted motions contain frequently occurring inter-human occlusions, which pose significant challenges to existing capturing algorithms. To address this problem, our key observation is that close physical interactions between two subjects typically happen under very specific situations (e.g., handshake, hug, etc.), and such situational contexts contain strong prior semantics to help infer the poses of occluded joints. In this spirit, we introduce reaction priors, which are invertible neural networks that bi-directionally model the pose probability distributions of one person given the pose of the other. The learned reaction priors are then incorporated into a query-based pose estimator, which is a decoder-only Transformer with self-attentions on both intra-joint and inter-joint relationships. We demonstrate that our design achieves considerably higher performance than previous methods on multiple benchmarks. What's more, as existing datasets lack sufficient cases of close human-human interactions, we also build a new dataset called Dual-Human to better evaluate different methods. Dual-Human contains around $2k$ *sequences of closely interacted two-person motions, each with synthetic multi-view renderings, contact annotations, and text descriptions. We believe that this new public dataset can significantly promote further research in this area. Our project page is at https://netease-gameai.github.io/Dual-Human/.*

1. Introduction

Despite the profound progress achieved in recovering human motions from monocular videos, two-person motion capture (MoCap), especially under close interactions, has rarely been addressed in the computer vision community. Currently, to tackle this task, one can either use single-person methods [11, 28, 32, 50] to separately obtain the motions of two actors, or use multi-person methods [39, 54, 55, 71] to simultaneously recover the poses of all seen people in the scene. Unfortunately, neither solution can robustly yield the desired results. Specifically, existing single-person methods are vulnerable to heavy occlusions since they do not take any interpersonal information into account, while multi-person methods concentrate much more on the correctness of the relative inter-person joint positions (*i.e.*, ordinal depth) than the precise pose interactions, though they typically exploit the situational contexts in some sense.

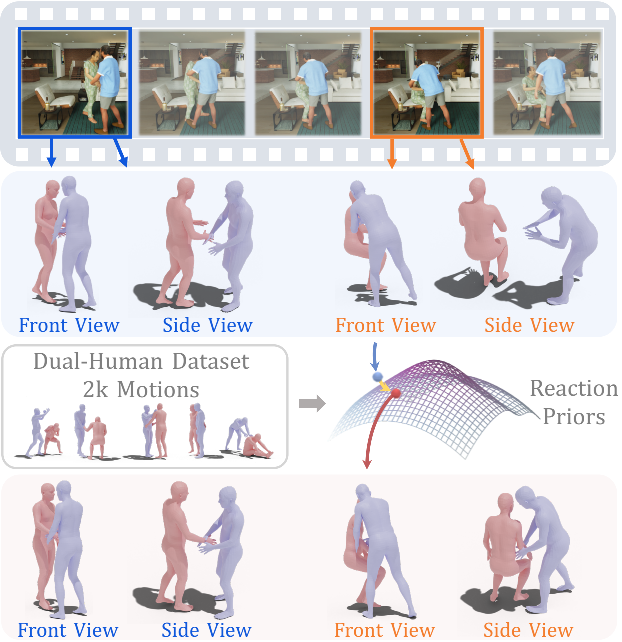


Figure 1. The proposed reaction priors can optimize the human motions from pose estimators for both less-occluded cases (blue box) and severely occluded cases (orange box).

Fortunately, recent research has proven that human-object interaction semantics show great success in assisting the pose prediction for both the occluded human and the occluded object [44, 62]. Likewise, prior information (*e.g.*, the actions and reactions between two closely interacted persons) should also have the potential to be leveraged for inferring the poses of occluded joints based on observed ones. Very recently, BUDDI [40] took the first step in this direction by learning static proxemic priors for 3D social interactions of two people from images via a diffusion model and illustrated the effectiveness of the priors in the downstream image-based human mesh recovery task.

In the same spirit, we also take advantage of priors, which we call reaction priors, to mitigate the ambiguity caused by occlusions (see Fig. 1). However, different from the diffusion-based single-frame proxemic priors in [40], which only depict the spatial relationships of two static poses, we build our reaction priors upon motion Variational AutoEncoder (VAE) and invertible neural networks (INN). Motion VAE is able to capture both the spatial and temporal information about the motion clips by mapping them into a unified latent space, while INN bi-directionally models the pose probability distributions of one person given the pose of the other. Note that the invertible design can elegantly address the interchangeable nature of two-person interactions. The reaction priors can be injected into pose estimators both as a training regularizer and an optimization guidance. We demonstrate that facilitated by the learned reaction priors, a simple query-based pose estimator (*i.e.*, a decoder-only Transformer network with self-attentions on both intra-joint and inter-joint relationships) is enough to outperform previous baselines on multiple benchmarks.

In addition, despite some recent advances in interacted motion datasets [12, 33, 65], a sizeable high-quality benchmark dataset dedicated to closely interacted motion capture remains absent in the community. To better evaluate the performance of various methods and to promote further research in this area, we build Dual-Human, a large-scale dataset containing around $2k$ sequences of closely interacted two-person motions, each with synthetic multi-view renderings, contact annotations, and text descriptions.

In summary, our key contributions are as follows:

- We introduce reaction priors to effectively model the pose probability distributions of one person given the pose of the interacted counterpart.
- We propose a new framework for capturing closely interacted two-person motions and demonstrate that by utilizing the learned reaction priors, this framework outperforms previous baselines on multiple benchmarks.
- We build Dual-Human, a large public dataset of closely interacted two-person motions with synthetic multi-view renderings, contact annotations, and text descriptions, which we believe will significantly promote further research in this area.

2. Related Work

In this section, we discuss related studies on human motion capture, reaction generation, and interacted human datasets.

Motion Capture: Though substantial progress has been made in MoCap from multi-view cameras [9, 10, 59, 70] or IMUs [14, 64], monocular methods cannot achieve stable performance due to depth ambiguity and occlusion. We discuss recent advances in three aspects: single-frame inputs, temporal inputs, and human priors.

For multi-person cases, single-frame methods can be divided into top-down and bottom-up. Top-down methods [39, 48] detect each person first and inherit single-person frameworks to estimate 3D poses. Bottom-up methods [13, 37, 71] infer intermediate representations from the whole image and group them into individuals. To simplify the multi-stage setting, one-stage works [53, 54, 60] leverage center and offset maps to perform one-shot inference.

Additionally, the temporal estimation task is also investigated. VIBE [27] uses GRUs and a discriminator to incorporate temporal information. Several recent approaches utilize Transformer-based structures to enhance temporal features [1, 41, 58]. 4D-Humans [16] adopts a vanilla Transformer decoder for each frame with an extra tracking module. Meanwhile, some other recent methods [55, 63, 66] focus more on the human global trajectory.

Another line of work explores human priors for optimization [4, 19, 43, 69]. VPoser [43] adopts the VAE structure with several regularization losses. ProHMR [29] uses the normalizing flow to construct image-conditioned priors. HuMoR [46] proposes the conditional prior via a conditional VAE. However, existing human priors are all relevant to a single person, while two-person priors also deserve explorations, especially in close interactions. InterPrior [73] is a recent work but only considers interacted hands. BUDDI [40] learns 3D proxemics priors via diffusion models for a static image. In contrast, we concentrate on the applications of the reaction priors in MoCap tasks.

Reaction Generation: Compared with other control signals [20, 21], reaction generation is to infer the reaction of one person from the action of the other person in interactions. Only a few works have been proposed until recently. Several papers approach this problem via various structural designs, *e.g.*, RNN-based [30] or Transformer-based [7] encoder-decoder. To prevent the decoder from creating monotonous reactions for all input actions, some researchers leverage the generative adversary network as a discriminator [15, 38]. However, these previous methods rely on accurate action conditions but MoCap estimations are usually noisy, making it difficult to apply these methods to MoCap directly.

Interacted Human Dataset: Most existing 3D multi-person datasets [2, 24, 36, 42] lack interacted motions because of the collection difficulty caused by close human-human distances. Early works use Kinect sensors [22,

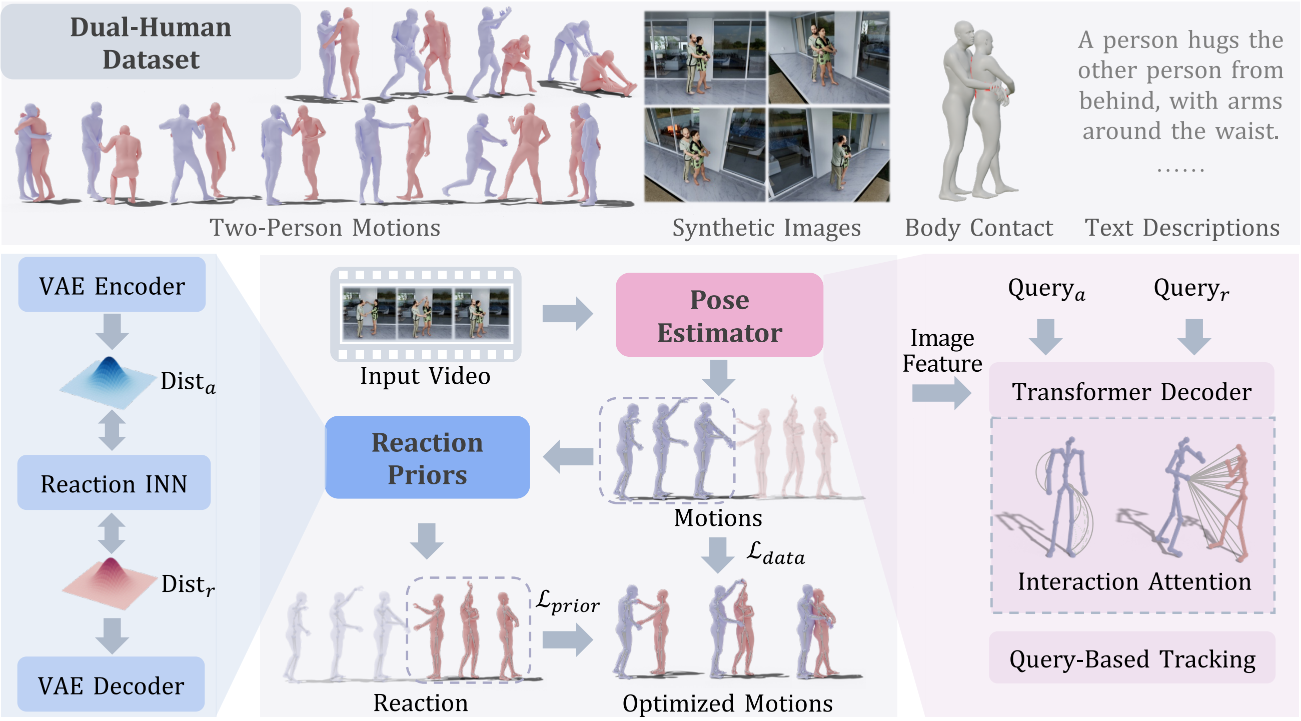


Figure 2. **Overview of Our Framework.** We build a large-scale richly annotated interacted human dataset, **Dual-Human**. Based on this dataset, we learn **reaction priors** composed of motion VAE and reaction INN, which bi-directionally model the pose probability distributions of one person given the pose of the other. The learned reaction priors are then incorporated into a query-based **pose estimator** with interaction-aware self-attention to optimize the motions.

26, 35, 56, 67] and thus have limited motion accuracy disturbed by sensor noises. 3DPW [57] is captured by IMUs but places more emphasis on single-person or less-interacted two-person motions. MultiHuman [72] is captured by sparse multi-view cameras, which hinder the accuracy in close interactions. CHI3D [12] provides additional contact annotations, but the types of interactions are relatively not enough. ExPI [18] is a more accurate dataset of extreme poses and does not cover daily interactions. Hi4D [65] is a very relevant dataset to us with high precision. However, Hi4D only has 100 motions because of the expensive collection manner. Inter-Human [33] is a recent dataset containing large-scale motions for text-to-motion tasks, but lacks images for MoCap. More importantly, its accuracy is influenced by the inability of multi-view RGB systems to solve severe occlusion issues. In contrast to them, our dataset tries to take accuracy, diversity, scale, and annotation completeness into consideration simultaneously.

3. Methods

In this section, we explain the pipeline of our work in detail, as shown in Fig. 2. We first discuss the concept of reaction priors and introduce how we statistically model the reaction priors using motion VAE and INN (Sec. 3.1). Then, we present our two-person MoCap framework which is built upon a simple Transformer-based pose estimator and the learned reaction priors (Sec. 3.2). After that, we give details of our Dual-Human dataset and its advantages over existing datasets (Sec. 3.3). Finally, we respectively describe the loss functions (Sec. 3.4) as well as implementation details (Sec. 3.5).

3.1. Reaction Priors

The reaction priors aim to infer the motion of the occluded person (*i.e.*, reaction) from the motion of the less-occluded person (*i.e.*, action), capturing the underlying dynamic rhythm and semantics in interactions, which is a kind of motion matching [5] in a sense. Normally, the motion is represented by 3D joint positions, rotations, and velocities. However, learning the motion directly may ignore the constraints of joint angle limits, thus producing unreasonable motions. Therefore, inspired by single-person priors [43, 46], we adopt the encoder-decoder structure. The initial step involves encoding the action into a compact latent representation in the form of mean and variance. This action latent representation is then utilized to generate the reaction latent representation. Since exchanging the action and reaction does not change the semantics of motions, we

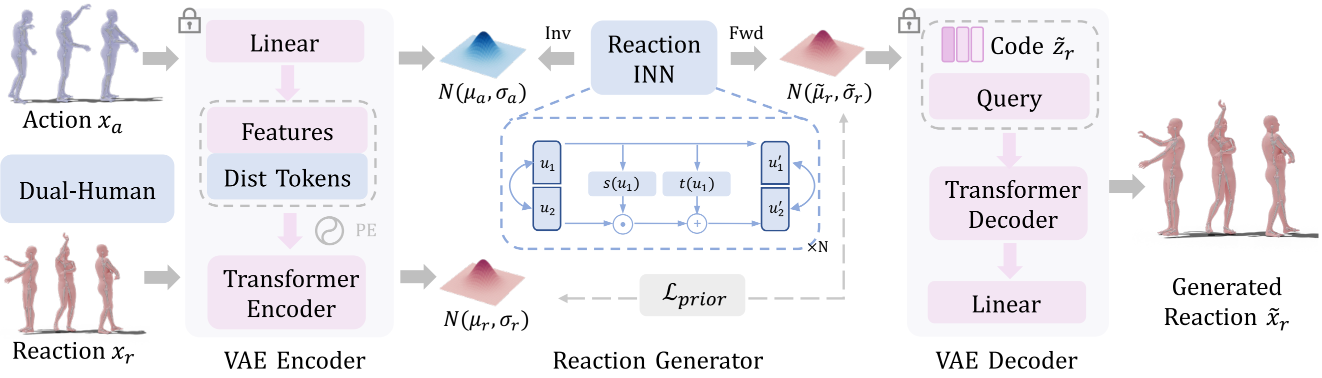


Figure 3. **Reaction Priors.** The reaction priors are composed of a VAE encoder, a reaction generator, and a VAE decoder. Initially, the VAE encoder maps the motion to the latent distribution. After that, the action distribution $\mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a)$ is fed to the reaction generator and converted to the reaction distribution $\mathcal{N}(\tilde{\boldsymbol{\mu}}_r, \tilde{\boldsymbol{\sigma}}_r)$, which are then recovered to the generated reaction.

leverage INN to construct our reaction generator to model the symmetry of human interactions. The invertibility of INN ensures that actions generate reactions and vice versa [31]. Finally, the reaction latent representation is decoded to the reaction.

Latent Motion Representation: We adopt the 269-dim motion representation for a person and the first 263-dim is the same as [17] including angular velocity, linear velocity, and height of root, as well as local joint positions, joint velocity, joint rotations, and foot contact. The last 6-dim includes initial direction and translation, which are only used to recover the camera-view motions. To make the motion representation compact, we leverage the VAE structure following [6, 45], as shown in Fig. 3. Specifically, the input motion $\boldsymbol{x}$ is mapped to the latent features, which together with distribution tokens, are fed to the Transformer encoder for predicting latent distribution parameters. The parameters include mean $\boldsymbol{\mu}$, log variance $\log \boldsymbol{\sigma}^2$ of the normal distribution. After that, a code $\boldsymbol{z}$ is sampled from the distribution and fed to the Transformer decoder to reconstruct the motion. The ELBO of VAE for optimization is as follows:

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q}[\log p(\boldsymbol{x}|\boldsymbol{z})] - \mathcal{D}_{KL}[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})]. \tag{1}$$

Reaction Generator: We utilize the reaction generator to infer the reaction latent distribution from the action latent distribution. Note that the subscripts 'a' and 'r' denote 'action' and 'reaction', respectively, and the tilde symbol indicates the corresponding generated variables. Suppose the joint distribution of two latent codes $p(\boldsymbol{z}_a, \boldsymbol{z}_r)$ follows the normal distribution. According to the properties of the normal distribution, the conditional probability $p(\boldsymbol{z}_r|\boldsymbol{z}_a)$ also follows the normal distribution. The probability of the generated reaction $\tilde{\boldsymbol{x}}_r$ conditioned on the action $\boldsymbol{x}_a$ can be approximated as follows via eliminating the direct dependence of $\tilde{\boldsymbol{z}}_r$ and $\tilde{\boldsymbol{x}}_r$ on the action $\boldsymbol{x}_a$ in that the action latent code $\boldsymbol{z}_a$ is in the underlying semantic space of $\boldsymbol{x}_a$:

$$p(\tilde{\boldsymbol{x}}_r|\boldsymbol{x}_a) \approx \iint \underbrace{p(\boldsymbol{z}_a|\boldsymbol{x}_a)\cdot p(\tilde{\boldsymbol{z}}_r|\boldsymbol{z}_a)}_{\mathcal{N}(\tilde{\boldsymbol{\mu}}_r,\tilde{\boldsymbol{\sigma}}_r)} \cdot p(\tilde{\boldsymbol{x}}_r|\tilde{\boldsymbol{z}}_r)\mathrm{d}\boldsymbol{z}_a\mathrm{d}\tilde{\boldsymbol{z}}_r, \tag{2}$$

where $p(\tilde{\boldsymbol{x}}_r|\tilde{\boldsymbol{z}}_r)$ is the decoder. $p(\boldsymbol{z}_a|\boldsymbol{x}_a)$ and $p(\tilde{\boldsymbol{z}}_r|\boldsymbol{z}_a)$ denote the encoder and the generator, respectively, and both of them follow the normal distribution, therefore their product also follows the normal distribution $\mathcal{N}(\tilde{\boldsymbol{\mu}}_r, \tilde{\boldsymbol{\sigma}}_r)$, which can be regarded as a learned *Gaussian conditional prior*.

Furthermore, to model the interchangeable nature of two-person interactions, we adopt the invertible neural network as the generator, which enables the seamless role switching of two people:

$$p(\tilde{\boldsymbol{z}}_r|\boldsymbol{x}_a) = p(\boldsymbol{z}_a|\boldsymbol{x}_a)\cdot\prod_k|\det(\frac{\partial f_k}{\partial \boldsymbol{z}_k})|^{-1}, \tag{3}$$

where f_k is the k-th invertible layer in INN.

Additionally, INN is also scalable for a relaxation of the assumption that the conditional probability $p(\tilde{\boldsymbol{z}}_r|\boldsymbol{z}_a)$ follows a more complex distribution rather than the normal distribution since its normalizing flow extension [8, 25] has been proven to be a powerful tool for modeling complex distributions with the Monte Carlo approximation for the calculation of KL divergence [47, 49].

In practice, instead of sampling $\boldsymbol{z}_a$ and $\tilde{\boldsymbol{z}}_r$ from the distribution, we perform inference directly between explicit distribution parameters $(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a)$ and $(\tilde{\boldsymbol{\mu}}_r, \tilde{\boldsymbol{\sigma}}_r)$. Specifically, the procedure of reaction priors is shown in Fig. 3. The encoder and decoder are trained in advance and frozen, while the INN is trained on our Dual-Human dataset. The separate training of the latent motion representation and reaction generator enables the former to leverage the large-scale single-person motion data.

Applications of Reaction Priors: The proposed reaction priors can be applied to both training regularization and test-time optimization. During the training of a motion estimator, L_1 or L_2 loss is generally adopted to supervise the motion estimation. However, they have no guarantee of structural constraints since a reverse-bent arm may also have small errors. Fortunately, the reaction priors can regularize the estimations via the following prior loss:

$$\mathcal{L}_{prior} = \mathcal{D}_{KL}[p(\boldsymbol{z}_r|\hat{\boldsymbol{x}}_r) \; || \; p(\tilde{\boldsymbol{z}}_r|\hat{\boldsymbol{x}}_a)], \tag{4}$$

where $p(\boldsymbol{z}_r|\hat{\boldsymbol{x}}_r)$ is the latent distribution from the reaction input, while $p(\tilde{\boldsymbol{z}}_r|\hat{\boldsymbol{x}}_a)$ is the latent distribution inferred from the action input through the reaction generator.

During test-time, the reaction latent distribution parameters $(\boldsymbol{\mu}_r^{opt}, \boldsymbol{\sigma}_r^{opt})$ can be the variables to be optimized. The objective functions include a data term and a prior term:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}_r^{opt},\boldsymbol{\sigma}_r^{opt}) &= ||\mathcal{D}(\boldsymbol{\mu}_r^{opt}, \boldsymbol{\sigma}_r^{opt}) - \hat{\boldsymbol{x}}_r||_2^2 \\ &+ \lambda_{prior} \cdot \mathcal{D}_{KL}[\mathcal{N}(\boldsymbol{\mu}_r^{opt}, \boldsymbol{\sigma}_r^{opt}) \; || \; p(\tilde{\boldsymbol{z}}_r|\hat{\boldsymbol{x}}_a)],\end{aligned} \tag{5}$$

where $\mathcal{D}$ denotes the decoder. λ_{prior} is a weight scalar. The data term aligns the output motions to the noisy observations and can also be applied to other forms of observations like 2D poses or partial 3D poses, while the prior term pushes the optimization variables towards the distribution produced by reaction priors.

Importantly, it is inevitable that the estimations from the MoCap framework are noisy, therefore we cannot train the reaction priors directly with accurate motions as it may cause limited generalization performance. To simulate the MoCap errors, we sample several camera views in a circle around the two performers and calculate the joint visibility in the corresponding view. This allows us to quantitatively assess the probability of each joint being visible under certain motions. Consequently, we add random noises to human poses in the training according to the probability to augment the data.

3.2. Pose Estimator

The learned reaction priors can be incorporated into a pose estimator framework for Monocular MoCap. Currently, mainstream methods of monocular multi-person 3D pose estimation are indirect, including top-down and bottom-up. However, the former discards the interacted spatial context caused by cropping, while the latter may fail to produce structured estimations since the prior on the number of people has not been used. Therefore, we turn to an end-to-end query-based decoder-only Transformer structure, encouraged by its success on 2D [34, 61] and 3D pose estimation [16, 59] tasks.

Generally, the query-based paradigm constructs the same number of queries as the variables to be estimated and adopts the Transformer decoder to output the estimations after fusing image features. Fig. 4 shows the pipeline of the proposed 3D pose estimator for two people. Given an input image, the feature extractor outputs image features via a common backbone network [51]. The features, as well as different types of queries ('Pose', 'Transl' and 'Prob' for 3D human poses, translation, and human probability, respectively), are fed to a decoder-only Transformer structure with self-attention, cross-attention, and feed-forward network, followed by multilayer perceptrons (MLPs) to regress the corresponding variables. Furthermore, our two-person assumption enables us to set the number of each type of query to 2. Besides, to eliminate the influence of image size and focal length, we normalize the depth in the translation representation by the field of view (FOV). As for the human probability, it can be understood as the salience score, indicating the possibility of how conspicuous a person is. We will use this variable to determine the more visible person. By now, although feasible, such a vanilla structure still faces the problem of low attention efficiency.

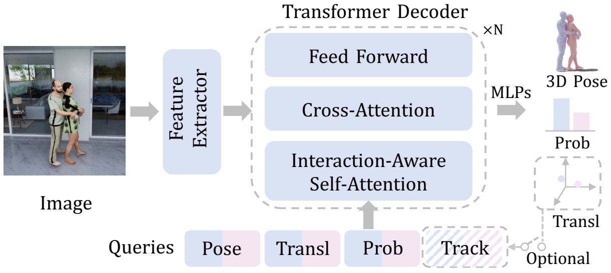


Figure 4. **3D Pose Estimator for Two People.**

Interaction-Aware Self-Attention: To improve the attention efficiency of the decoder, inspired by [7, 34], we propose an interaction-aware self-attention, which consists of intra-human self-attention and inter-human self-attention sequentially, as shown in Fig. 2. The queries of two people are separately attended in intra-human self-attention, whereas the 'Pose' queries only focus on their ancestors and children on the skeleton. For inter-human self-attention, 'Pose' and 'Transl' queries are attended mutually since any two joints of two people may interact. Besides, two 'Prob' queries are restricted to relating only to each other.

Furthermore, instead of requiring an additional tracking module for temporal inputs in common practice [16], we try to implement the tracking mechanism using queries [52].

Query-Based Tracking: Specifically, the translation estimations (root positions) of the previous frame can serve as the tracking information to guide the attention in the current frame to distinguish two people, thus a certain query will always focus on the same person. We concatenate these 'Track' queries with other queries as additional inputs.

After obtaining the tracked motions, we impose temporal smoothness and consistency via SmoothNet [68]. Then

Datasets	Sources	Motions	Subjects	Other Annotations		
				Image	Contact	Text
3DPW [57]	RGB+IMUs	29	8	✓		
CHI3D [12]	MV+Optical	626	5	✓	✓	
ExPI [18]	MV+Optical	115	2	✓		
Hi4D [65]	MV RGB+IR	100	40	✓	✓	
Inter-Human [33]	MV RGB	6022	-			✓
Dual-Human	IMUs+VIVE	2019	40	✓	✓	✓

Table 1. **Comparison of Human Interacted Datasets.** Only two-person motions are counted. 'MV' is multi-view. Inter-Human [33] is a very recent work for motion generation rather than capture, thus has limited accuracy, which is listed for completeness.

we further improve the performance through the proposed reaction priors. Specifically, we adopt a sliding window of length L with an overlap of $L/2$ to split the videos with various lengths. Within each window, we leverage reaction priors to optimize the estimations as mentioned before.

3.3. Dual-Human

Existing prevailing datasets for interacted people face the problems of being small in scale and limited diversity. Inter-Human [33] is large but is not designed for interacted MoCap because its covered motions have issues such as being interpenetrated, less-interacted, and relatively insufficient accuracy for MoCap. To address this problem, we build a large and easily scalable dataset called Dual-Human for learning the interactive situational contexts of two people. The comparison between Dual-Human and existing datasets is shown in Table 1. Our Dual-Human dataset has around $2k$ self-collected two-person motions, which is much more than previous relevant datasets for MoCap. Apart from human motions, we also provide synthetic multi-view images (see samples in Fig. 5), body contact, and text descriptions, supporting other tasks like text-to-motion. More details can be found in Sec. 4 and the supplementary material.

3.4. Objective Functions

To train our framework, multi-stage training strategies are adopted to take advantage of the properties of each module.
Pose Estimator: The pose estimator is supervised by three losses: $\mathcal{L}_{pose}$, $\mathcal{L}_{transl}$, and $\mathcal{L}_{prob}$. The smooth L_1 loss is applied to the 3D human poses. For the case without the tracking information, we use the Hungarian matching algorithm to match the estimations and the ground-truths. Besides, the L_2 loss is utilized for the translation while the cross-entropy loss is used for the human saliency probability.
Motion VAE: To train the motion VAE for latent representations, we adopt the common loss according to ELBO (Eq. 1) as follows:

$$\mathcal{L}_{vae} = ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_1 + \lambda_{dist} \cdot \mathcal{D}_{KL}[p(\boldsymbol{z}|\boldsymbol{x}) \, || \, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})], \quad (6)$$

where λ_{dist} is a weight scalar.

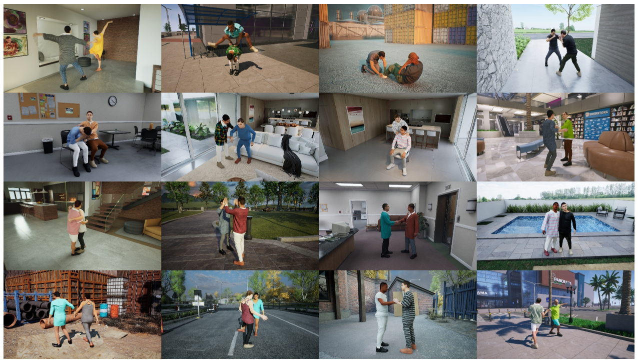
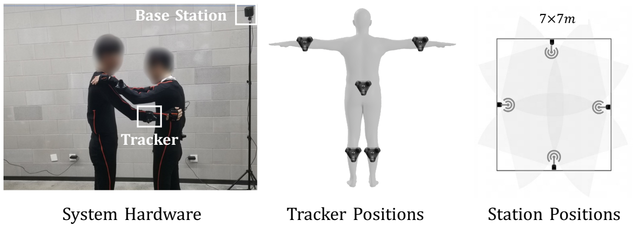


Figure 5. **Sample Images in Dual-Human.**



System Hardware Tracker Positions Station Positions

Figure 6. **System Configurations of Dual-Human.**

Reaction INN: With the frozen motion encoder and decoder, we train the reaction INN by supervising the generated reaction latent distribution in the same form as $\mathcal{L}_{prior}$ (Eq. 4), except that we use ground-truth actions and reactions with noise augmentation here.
SmoothNet: After transforming the estimated 3D poses to our 269-dim motion representation, we train SmoothNet with the L_1 motion loss.

3.5. Implementation Details

The input image of two people has a resolution of 512×512. We adopt HRNet-W48 [51] as the backbone of our pose estimator, which is trained for 300 epochs with an initial learning rate of 1×10^{-5}, decayed with a factor of 5. The motion VAE, reaction INN, and SmoothNet are trained for 4000, 500, and 2000 epochs, respectively, with a learning rate of 1×10^{-4}. The AdamW optimizer is applied to all networks with a weight decay of 5×10^{-4}. The weight scalars are set as follows: $\lambda_{pose} = 10$, $\lambda_{transl} = 1$, $\lambda_{prob} = 1$, $\lambda_{dist} = 0.0001$. Besides, λ_{prior} is adjusted flexibly in training or test-time optimization, and we set it to 0.01 after comparisons. The window size L is set to 64.

4. Constructing Dual-Human Dataset

In this section, we introduce how to construct our Dual-Human dataset. Specifically, we adopt an inertial MoCap system, consisting of Xsens suits and several HTC Vive devices that can alleviate drift issues and improve the posi-

Methods	Hi4D			CHI3D			Dual-Human		
	MPJPE ↓	PA ↓	Transl ↓	MPJPE ↓	PA ↓	Transl ↓	MPJPE ↓	PA ↓	Transl ↓
CLIFF [32] (Top-Down)	77.9	60.4	285.8	76.4	**45.9**	375.9	73.3	53.1	275.2
4D-Humans [16] (Transformer)	80.7	62.2	-	73.1	48.6	-	66.9	51.3	-
BEV [54] (Bottom-Up)	89.2	**59.3**	212.0	89.1	54.6	314.1	83.1	56.8	263.2
TRACE [55] (Temporal)	83.8	60.4	179.9	75.9	49.7	236.3	67.8	53.9	120.9
Ours	**75.0**	59.7	**106.7**	**71.9**	47.8	**228.5**	**63.4**	**51.2**	**112.1**

Table 2. **Results on Interacted Human Benchmarks.** 'PA' is PA-MPJPE.

tioning ability, as shown in Fig. 6. This system can capture large-scale relatively accurate interacted human motions.

After collecting original motions, we fit the parametric human model SMPL-X [43] for a unified representation, which can be formulated as an energy minimization problem over body shape, pose, and translation parameters. However, penetration is inevitable since the inertial MoCap system cannot capture the surface of the human body. Therefore, we add a collision loss based on Signed Distance Field (SDF) following [23] and jointly optimize the parameters of interacted people. Besides, We determine a vertex is in contact if its SDF value is below a certain threshold and automatically extract vertex-level contact annotations, which are crucial for exploring close human interactions.

As for the corresponding RGB images, we follow the procedure of BEDLAM [3] to render realistic multi-view images with various textures and cloth simulation. Apart from these, the text descriptions are manually annotated and provided, which can be utilized in generative tasks.

5. Experiments

In this section, we validate the effectiveness of our framework on several benchmarks compared with previous approaches, demonstrate key designs via an ablation study, and discuss the limitations and future work.

5.1. Experimental Setup

We perform experiments on the following datasets:
Hi4D [65] is an indoor dataset with a small amount of but accurate motions. We pick $1/8$ of all sequences as the test set covering different types of interactions and the remaining as the training set.
CHI3D [12] is an indoor dataset with 8 interaction types. We use subjects $(02, 04)$ as the training set and evaluate on subject 03 since the former two have the same performer, which can prevent overfitting to a single appearance.
Dual-Human covers both indoor and outdoor scenes with various interactions and textures. We split the dataset into a training set and a test set with a ratio of $3:1$.
Metrics include MPJPE (mm), PA-MPJPE (mm), and Translation Error (mm). MPJPE measures the accuracy of the 3D root-relative pose. It calculates the distance between the predicted and the ground-truth joint locations averaged over all joints. PA-MPJPE is similar but performs a rigid alignment before MPJPE. Translation Error is defined as the root joint error measured by Euclidean distance.

Priors	MPJPE ↓	PA ↓	Transl ↓
Proxemics Priors (BUDDI) [40]	67.3	53.4	126.1
Reaction Priors (Ours)	**64.1**	**50.3**	**115.2**

Table 3. **Comparisons of Two-Person Priors.**

5.2. Comparison with Previous Methods

We perform thorough experiments comparing our method to baselines using the same setting on several benchmarks. The baseline methods include the top-down method CLIFF [32], the bottom-up method BEV [54], the transformer-based method 4D-Humans [16], and the temporal method TRACE [55]. Table 2 shows that our method outperforms these baselines in most metrics on three benchmarks. Besides, we also compare the effectiveness of priors in Table 3. We use the same pose estimator and optimize the results through BUDDI [40] and our reaction priors, respectively. As the optimization of BUDDI is time-consuming, we only evaluate on a 10% randomly picked test set. The comparisons show that our priors achieve better performance. Fig. 7 shows the qualitative comparisons, revealing that existing algorithms may produce inaccurate human positions, unreasonable interaction poses, and missing people when severe occlusions or close interactions occur, while our method can alleviate these problems to some extent.

5.3. Ablation Study

To validate our key designs, we conduct comparative experiments on reaction priors and MoCap, respectively, as shown in Table 4.

For reaction priors, we design the following baselines: (a) mapping actions to reactions directly without using motion VAE; (b) replacing the INN with MLPs; (c) using clean actions without noise augmentation in training. It can be observed that these changes will degrade the performance. First, if motion VAE is discarded, the errors will increase possibly due to unreasonable motions. Second, INN will converge faster and better than MLPs due to its

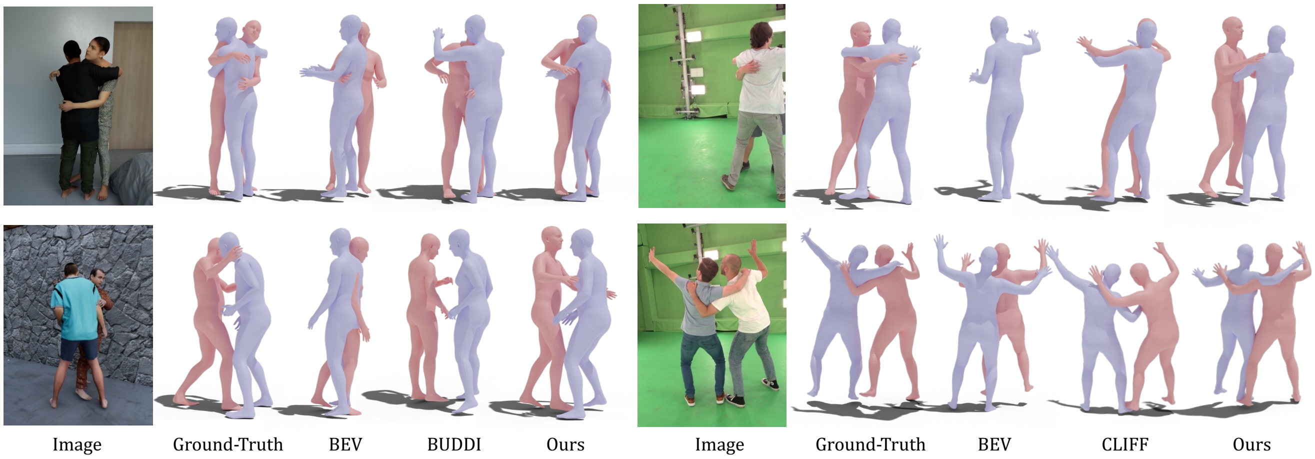


Figure 7. **Qualitative Comparisons.** We compare our approach with previous methods including BEV [54], BUDDI [40] and CLIFF [32] on Dual-Human and Hi4D. The best view that can show the differences is chosen for rendering.

Designs		Dual-Human		
		MPJPE ↓	PA ↓	Transl ↓
Reaction Generation	(a) W/o motion VAE	64.4	52.8	115.7
	(b) Replace INN with MLPs	66.2	53.1	127.5
	(c) W/o noise augmentation	77.4	60.5	141.3
MoCap	(a) W/o interaction attention	70.3	55.9	125.5
	(b) W/o reaction priors	67.9	54.2	118.2
	(c) W/o temporal network	63.6	51.8	128.3
	(d) Ours (full model)	**63.4**	**51.2**	**112.1**

Table 4. **Ablation Study of Reaction Generation and MoCap.**

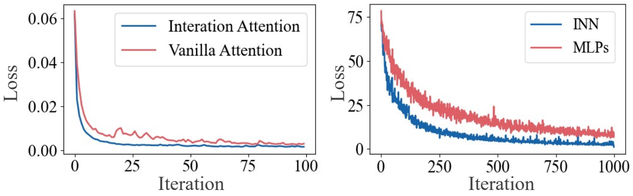


Figure 8. **Convergence Comparisons.** Left: interaction-aware self-attention v.s. the vanilla attention for the pose estimator. Right: INN v.s. MLPs for the reaction generator.

bi-directional nature, as shown in Fig. 8. Third, if we do not add noises during training, significant degradation will occur in testing, because the priors cannot generalize to the noisy raw estimations from the pose estimator well.

For the MoCap framework, we validate the effectiveness of three components: interaction-aware self-attention, reaction priors, and temporal network. The interaction-aware self-attention is crucial for the pose estimator as it not only enhance the performance but also accelerates the convergence, as shown in Fig. 8. It can also be observed that the reaction priors improve both the pose and translation accuracy, and the improvement is more significant for motions that have severe occlusions. Besides, the temporal network (SmoothNet we used) mainly reduces the translation error.

5.4. Limitation and Future Work

Our method still has several limitations. First, our reaction priors cannot handle interactions that are very different from the motions in our dataset. Furthermore, for input cases with no interactions, very large values of λ_{prior} are likely to produce false interaction motions. Second, despite that we focus on two-person close interactions, our assumption that a scene must have two people is relatively limited, requiring additional processes for more than two people.

For future work, we will further expand the scale and variety of Dual-Human to cover more interactions. Notice that we also provide text descriptions for the motions in Dual-Human, which means this dataset may also be potentially used for research on multi-modal (*e.g.*, text, image, *etc.*) interacted motion generation.

6. Conclusion

In this work, we attack a challenging problem, *i.e.*, capturing closely interacted two-person motions from monocular videos. Specifically, we introduce reaction priors to quantitatively describe the probability distribution of one person's pose conditioned on the other's, and propose an effective two-person MoCap framework by incorporating the learned reaction priors with a decoder-only Transformer-based human pose estimator. Apart from the technical contributions, we also build Dual-Human, a large-scale high-quality dataset on closely interacted two-person motions that contains $2k$ richly annotated (multi-view renderings, contact annotations, and text descriptions) motions. We believe our work provides a sound foundation for further research on human-human interaction capture.

References

[1] Fabien Baradel, Romain Brégier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez. Posebert: A generic transformer module for temporal 3d human modeling. *IEEE TPAMI*, 2022. 2

[2] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, pages 1669–1676, 2014. 2

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 7

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578, 2016. 2

[5] Michael Büttner and Simon Clavet. Motion matching and the road to next-gen animation, 2015. 3

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2023. 4

[7] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE TMM*, 2023. 2, 5

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *ICLR*, 2016. 4

[9] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE TPAMI*, 44(10):6981–6992, 2021. 2

[10] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, pages 12814–12823, 2021. 2

[11] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *CVPR*, pages 8781–8791, 2023. 1

[12] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 2, 3, 6, 7

[13] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *NeurIPS*, 34:19385–19397, 2021. 2

[14] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *IJCV*, 127(4):381–397, 2019. 2

[15] Aman Goel, Qianhui Men, and Edmond SL Ho. Interaction mix and match: Synthesizing close interaction using conditional hierarchical gan with multi-hot class embedding. In *Comput. Graph. Forum*, pages 327–338, 2022. 2

[16] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 5, 7

[17] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 4

[18] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *CVPR*, pages 13053–13064, 2022. 3, 6

[19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019. 2

[20] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM TOG*, 39(6):1–14, 2020. 2

[21] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM TOG*, 39(4): 53–1, 2020. 2

[22] Tao Hu, Xinyan Zhu, Wei Guo, Kehua Su, et al. Efficient interaction recognition through positive action representation. *Mathematical Problems in Engineering*, 2013, 2013. 2

[23] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 7

[24] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 2017. 2

[25] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *NeurIPS*, 31, 2018. 4

[26] Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. *IJRR*, 40(4-5):691–697, 2021. 3

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2

[28] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 1

[29] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, pages 11605–11614, 2021. 2

[30] Jogendra Nath Kundu, Himanshu Buckchash, Priyanka Mandikal, Anirudh Jamkhandi, Venkatesh Babu Radhakrishnan, et al. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *WACV*, pages 2724–2733, 2020. 2

[31] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *CVPR*, pages 12933–12942, 2023. 4

[32] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pages 590–606, 2022. 1, 7, 8

[33] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv*, 2023. 2, 3, 6

[34] Huan Liu, Qiang Chen, Zichang Tan, Jiangjiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, Yao Zhao, and Jingdong Wang. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *ICCV*, 2023. 5

[35] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI*, 42(10):2684–2701, 2019. 3

[36] 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. In *3DV*, 2018. 2

[37] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM TOG*, 39(4): 82–1, 2020. 2

[38] Qianhui Men, Hubert PH Shum, Edmond SL Ho, and Howard Leung. Gan-based reactive motion synthesis with class-aware discriminators for human–human interaction. *Computers & Graphics*, 102:634–645, 2022. 2

[39] 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. In *ICCV*, pages 10133–10142, 2019. 1, 2

[40] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 2, 7, 8

[41] 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 *ICCV*, pages 14772–14782, 2023. 2

[42] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 2

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2, 3, 7

[44] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *CVPR*, pages 4726–4736, 2023. 1

[45] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 4

[46] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, pages 11488–11499, 2021. 2, 3

[47] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015. 4

[48] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE TPAMI*, 2019. 2

[49] Hendra Setiawan, Matthias Sperber, Udhay Nallasamy, and Matthias Paulik. Variational neural machine translation with normalizing flows. In *ACL*, 2020. 4

[50] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation. In *CVPR*, pages 574–584, 2023. 1

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 5, 6

[52] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv*, 2020. 5

[53] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 2

[54] 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 *CVPR*, pages 13243–13252, 2022. 1, 2, 7, 8

[55] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *CVPR*, pages 8856–8866, 2023. 1, 2, 7

[56] Coert Van Gemeren, Ronald Poppe, and Remco C Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In *International Workshop on Human Behavior Understanding*, pages 116–133, 2016. 3

[57] 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 *ECCV*, pages 601–617, 2018. 3, 6

[58] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *ICCV*, pages 13033–13042, 2021. 2

[59] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d pose estimation. *NeurIPS*, 34:13153–13164, 2021. 2, 5

[60] Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, and Si Liu. Distribution-aware single-stage models for multi-person 3d pose estimation. In *CVPR*, pages 13096–13105, 2022. 2

[61] Yabo Xiao, Kai Su, Xiaojuan Wang, Dongdong Yu, Lei Jin, Mingshu He, and Zehuan Yuan. Querypose: Sparse multi-person pose regression via spatial-aware part-level query. *NeurIPS*, 35:12464–12477, 2022. 5

[62] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *CVPR*, pages 4757–4768, 2023. 1

[63] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, pages 21222–21232, 2023. 2

[64] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *CVPR*, pages 13167–13178, 2022. 2

[65] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, 2023. 2, 3, 6, 7

[66] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, pages 11038–11049, 2022. 2

[67] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPRW*, pages 28–35, 2012. 3

[68] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *ECCV*, pages 625–642. Springer, 2022. 5

[69] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11343–11353, 2021. 2

[70] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, pages 1324–1333, 2020. 2

[71] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, pages 550–566, 2020. 1, 2

[72] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *ICCV*, pages 6239–6249, 2021. 3

[73] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, pages 9054–9064, 2023. 2