

# Reconstructing Close Human Interactions from Multiple Views

QING SHUAI\*, State Key Laboratory of CAD&CG, Zhejiang University, China

ZHIYUAN YU\*, Hong Kong University of Science and Technology, China

ZHIZE ZHOU, Capital University of Physical Education and Sports, China

LIXIN FAN and HAIJUN YANG, WeBank, China

CAN YANG, Hong Kong University of Science and Technology, China

XIAOWEI ZHOU†, State Key Laboratory of CAD&CG, Zhejiang University, China

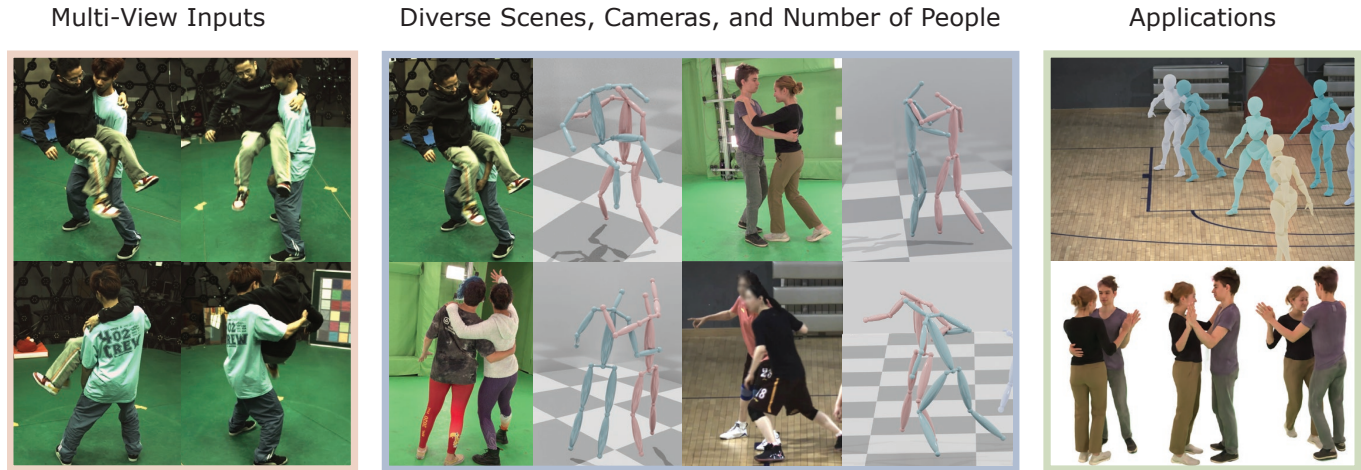


Fig. 1. Our system is designed to recover the 3D poses of individuals engaging in close-range interactions, utilizing input from multiple calibrated cameras. We introduce a novel learning-based approach that effectively handles occlusions and interactions between individuals at close quarters. The standout feature of our system, which allows it to be trained without real data, enables the system to handle various scenes, camera configurations, and number of individuals. Our system facilitates a broad range of real applications, such as character animation (top-right) and free-viewpoint video synthesis (bottom-right).

This paper addresses the challenging task of reconstructing the poses of multiple individuals engaged in close interactions, captured by multiple calibrated cameras. The difficulty arises from the noisy or false 2D keypoint detections due to inter-person occlusion, the heavy ambiguity in associating keypoints to individuals due to the close interactions, and the scarcity of

\*The first two authors contributed equally to this work.

†Corresponding author: Xiaowei Zhou.

Authors' addresses: Qing Shuai, s\_q@zju.edu.cn, State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China; Zhiyuan Yu, zyuaq@ust.hk, Hong Kong University of Science and Technology, Mathematics, Hong Kong, China; Zhize Zhou, zhouzhize@cupes.edu.cn, Capital University of Physical Education and Sports, Beijing, China; Lixin Fan, lixinfan@webank.com; Haijun Yang, navyyang@webank.com, WeBank, Shenzhen, China; Can Yang, macyang@ust.hk, Hong Kong University of Science and Technology, Mathematics, Hong Kong, China; Xiaowei Zhou, xwzhou@zju.edu.cn, State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

0730-0301/2023/12-ART273 \$15.00

<https://doi.org/10.1145/3618336>

training data as collecting and annotating motion data in crowded scenes is resource-intensive. We introduce a novel system to address these challenges. Our system integrates a learning-based pose estimation component and its corresponding training and inference strategies. The pose estimation component takes multi-view 2D keypoint heatmaps as input and reconstructs the pose of each individual using a 3D conditional volumetric network. As the network doesn't need images as input, we can leverage known camera parameters from test scenes and a large quantity of existing motion capture data to synthesize massive training data that mimics the real data distribution in test scenes. Extensive experiments demonstrate that our approach significantly surpasses previous approaches in terms of pose accuracy and is generalizable across various camera setups and population sizes. The code is available on our project page: <https://github.com/zju3dv/CloseMoCap>.

CCS Concepts: • **Computing methodologies** → **Motion capture**.

Additional Key Words and Phrases: human pose estimation, motion capture

## ACM Reference Format:

Qing Shuai, Zhiyuan Yu, Zhize Zhou, Lixin Fan, Haijun Yang, Can Yang, and Xiaowei Zhou. 2023. Reconstructing Close Human Interactions from Multiple Views. *ACM Trans. Graph.* 42, 6, Article 273 (December 2023), 14 pages. <https://doi.org/10.1145/3618336>

## 1 INTRODUCTION

Markerless human motion capture is a vital enabling technology with broad applications spanning character animation, motion analysis, and 3D reconstruction of dynamic events. Compared to active or passive motion capture techniques, it has lower hardware requirements and fewer environmental constraints, capable of capturing human movement from a single or multiple calibrated cameras. This makes it more adaptable to a wide array of downstream needs.

In this paper, our primary focus is the estimation of multi-person 3D poses from multiple calibrated cameras, particularly in scenarios involving close-range interactions. This is critical in many real scenarios, especially those that involve human-human interactions, such as multi-person dance performances and basketball games. These close-range interactions cause substantial challenges to pose estimation methods due to severe occlusions and heavy ambiguities when detecting and associating human body keypoints in multi-view images. Traditional methods [Dong et al. 2019; Zhang et al. 2020; Zhou et al. 2022] typically first estimate 2D poses from each viewpoint, then associate different 2D instances or keypoints across these views, and finally estimate 3D poses through triangulation. However, such association-based methods suffer from 2D estimation errors, especially in cases of close-range interaction. Additionally, the process of keypoint association becomes extremely challenging in crowded scenarios. This often results in either missed keypoints or low-precision pose estimates for these methods.

In contrast, learning-based approaches [Tu et al. 2020; Wang et al. 2021; Wu et al. 2021; Ye et al. 2022] avoid performing 2D association. Instead of using keypoints as input directly, they construct feature volumes from multi-view 2D feature maps and then regress poses directly in 3D space in an end-to-end fashion. However, these learning-based methods heavily rely on paired 2D-3D ground-truth data for training. Existing datasets are typically acquired indoors using marker-based methods [Ionescu et al. 2014] or dense multi-view triangulation [Joo et al. 2015] for ground-truth annotation. These datasets often lack diversity in terms of performers, actions, camera configurations, and background scenes. They also struggle to capture and annotate complex interactions due to severe occlusions. As a result, learning-based methods trained on these datasets are hard to generalize to different scenarios involving close interactions (as shown in Fig. 3). The change in camera placement can significantly impact the outcomes, thereby limiting the generalization capability of learning-based methods in real-world applications.

To overcome these challenges, we propose a novel system that is designed for close interaction scenarios and can be trained with only synthetic data. The proposed method initially estimates keypoint heatmaps from multi-view images. Subsequently, it reconstructs the 3D centers for each individual. These centers are used to construct feature volumes, which are then passed through the 3D volumetric network. Finally, the network outputs the 3D pose of each individual. Previous approaches often directly regress the 3D pose of each individual using the keypoint feature volume derived from 2D keypoint heatmaps. However, these methods struggle to represent close interactions among multiple individuals. When two individuals are in close proximity (as shown in Fig. 2), their pelvis-centered feature volumes are highly similar, which may create ambiguity

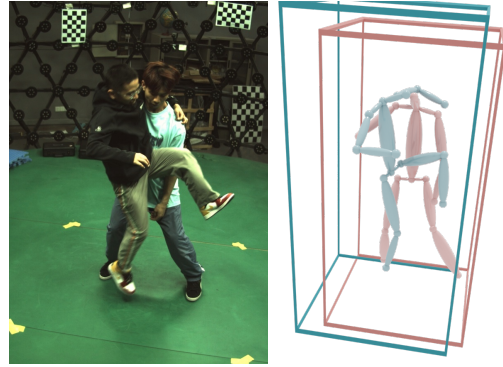


Fig. 2. **Challenges in pose estimation with close proximity.** This image highlights that when two individuals are in close proximity, it becomes difficult to obtain accurate 2D pose estimates due to heavy inter-person occlusion and keypoint association ambiguity. Moreover, in learning-based methods that directly regress 3D poses from feature volumes, the similarity in constructed volumes due to their spatial closeness complicates keypoint distinction for regression networks.

in the network output. To address this issue, we incorporate the estimated 3D centers for all individuals as an additional input to the network. This information is used to build the anchor-guided feature volumes, serving as a conditional signal.

Typically, image-based pose regression networks demand substantial quantities of multi-view image-3D pose pairs for training, which are often challenging to obtain. Fortunately, our method relies solely on 2D keypoints heatmaps as input. With known 3D poses and camera parameters, we can synthesize multi-view 2D heatmaps to train our network. For common multi-view tasks, we leverage existing camera parameters and a wealth of MoCap data to generate these 2D heatmaps. To enhance realism, we apply data augmentation to closely mimic heatmaps from real close interaction scenarios. The vast MoCap dataset provides a rich set of motion data, improving network robustness. Besides, 2D heatmaps can be easily obtained during inference with off-the-shelf 2D pose estimators, making our method user-friendly and practical.

We conduct experiments on the latest close interaction datasets, which demonstrate that our method significantly outperforms previous approaches in terms of accuracy and robustness. We also validate our method across various scenarios with different scene scales, number of people, and camera configurations (see Fig. 1), affirming its robustness and applicability.

Our contributions are summarized as follows:

- We propose a novel system to solve the problem of multi-person markerless motion capture in close interaction scenarios.
- We tackle the challenge of training data scarcity by generating synthetic samples using known camera parameters and extensive MoCap data.
- Through comprehensive experiments, we demonstrate our method's superior performance and applicability across various challenging scenarios.

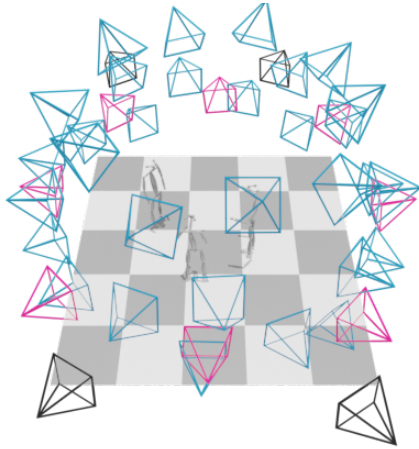


Fig. 3. **Diverse camera position and orientation in various datasets.** We show the camera positions and orientations of several common datasets: Panoptic [Joo et al. 2015] (blue), CH13D [Fieraru et al. 2020] (black), and Hi4D [Yin et al. 2023] (red). These datasets, collected by different studios, exhibit significant variations. In practical scenarios like a basketball court, scale differences become more apparent, posing challenges for network generalization across datasets.

## 2 RELATED WORKS

*Multi-view multi-person 3D pose estimation.* Given calibrated multi-view images, existing approaches can be divided into two main categories: association-based methods and learning-based methods. Association-based methods usually take associated 2D poses as input and lift them into 3D poses. While learning-based methods usually associate and regress 3D poses from 2D keypoint heatmaps or detections with deep neural networks.

Association-based methods usually adopt a two-stage approach. In the first stage, 2D human poses are independently estimated for each view using pre-trained models [Cao et al. 2017; Wang et al. 2020b]. In the second stage, the estimated 2D poses of the same individual in different views are associated and lifted to 3D coordinates via triangulation [Dong et al. 2019; Huang et al. 2020; Zhou et al. 2022], plane sweep stereo [Lin and Lee 2021], or the pictorial model [Belagiannis et al. 2014, 2015]. Existing methods mainly differ in the second stage. Dong et al. [2019] grouped the 2D instances from geometric and appearance cues. Huang et al. [2020] proposed a novel end-to-end training scheme including a dynamic matching algorithm for association. Zhang et al. [2020] proposed a 4D association method that first built a 4D graph from sequential 2D poses and applied a bundle Kruskal’s algorithm to search and assemble limbs based on the 4D graph. Lin and Lee [2021] associated multi-view 2D poses based on plane sweep stereo and then applied 1D convolution neural networks (CNNs) to regress keypoints depth for 3D pose estimation. Zhou et al. [2022] considered all plausible skeletons via a tree-structure graph and reformulated the association problem as mode seeking. However, it is hard for these methods to deal with missing or erroneous 2D keypoints.

Learning-based methods typically take 2D keypoints heatmaps or feature maps as input and convert them into 3D features. These

3D features are then used to regress multiple 3D poses directly using deep neural networks. As a pioneer, VoxelPose [Tu et al. 2020] extended the learnable single-person pose estimator [Iskakov et al. 2019] to multi-person scenarios. It first built 3D score volumes from multi-view 2D heatmaps. Then it used 3D CNNs to localize each individual from a coarse volume and estimate their 3D poses from fine volumes. The follow-up work primarily enhanced VoxelPose in terms of speed and performance. Faster-VoxelPose [Ye et al. 2022] realized real-time pose estimation by using BEV representation for localization and replacing computationally expensive 3D CNNs with efficient tri-plane 2D CNNs. Meanwhile, Wu et al. [2021] applied graph convolution networks (GCNs) for human localization and pose regression, achieving significant improvements in terms of performance and computation complexity. Taking a different approach, Wang et al. [2021] proposed a multi-view Transformer to predict multi-person 3D joint positions directly from multi-view images without the need for human localization. Although these methods reduce the reliance on 2D pose estimation, they do not consider close human interactions.

*Single-view multi-person 3D pose estimation.* Compared to multi-view inputs, single-view pose estimation presents a more accessible option, leading many methods to focus on directly estimating the poses of multiple people from a single image. These methods can be primarily categorized into two classes: top-down and bottom-up. Top-down methods first perform human detection and then estimate 3D poses for each detected individual. Moon et al. [2019] first estimate the root depth of the human body and then the root-relative 3D pose. Subsequent works [Benzine et al. 2020; Lin and Lee 2020; Wang et al. 2020a, 2022] improve estimation accuracy by considering factors like occlusion, depth relations, and joint distribution. Bottom-up methods [Mehta et al. 2020; Zhen et al. 2020] first predict the 3D locations of all human keypoints and then associate them with each individual. Instead of 3D skeleton representation, more recent works [Cha et al. 2022; Fieraru et al. 2021c; Qiu et al. 2023; Sun et al. 2022; Ye et al. 2023; Yuan et al. 2022] have focused on recovering multiple SMPL [Loper et al. 2015] or GHUM [Xu et al. 2020] meshes from monocular images or videos. Despite the remarkable progress in monocular 3D human pose estimation, these techniques still suffer from depth ambiguity and occlusion, making it difficult for them to obtain high-precision estimation.

*3D human pose datasets.* Currently there exist many datasets providing RGB images and 3D pose annotations in both single-person and multi-person scenarios. Given the ease of data collection, there are numerous single-person 2D [Joo et al. 2021; Kolotouros et al. 2019; Lin et al. 2014] and 3D datasets [Cai et al. 2022; Chatzitofis et al. 2020; Fieraru et al. 2021a,b; Ionescu et al. 2014; Mehta et al. 2017; Ofli et al. 2013; Sigal et al. 2010; Trumble et al. 2017; Von Marcard et al. 2018; Yoon et al. 2021], featuring diverse actors, actions, and modalities. However, training and validation data are relatively scarce in multi-person scenarios. Commonly used datasets such as Shelf and Campus [Belagiannis et al. 2014] contain a limited number of frames, scenarios, and actions, making them less appropriate for comprehensive training and evaluation. Panoptic [Joo et al. 2015], currently the largest real-world multi-person dataset, mainly focuses



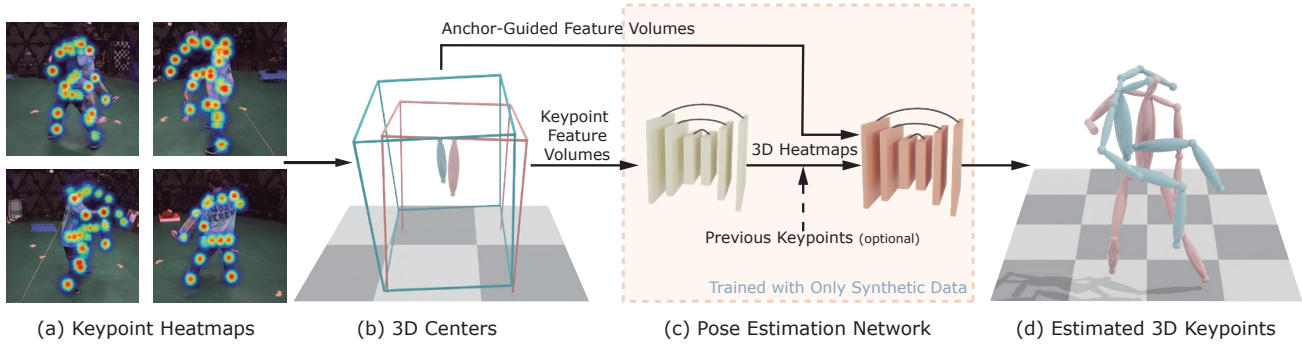


Fig. 4. **Illustration of our method.** For a multi-view scene, we first estimate (a) the 2D keypoint heatmaps of all people from input images. We then recover (b) the 3D centers of all people from these heatmaps. Following this, we construct keypoint feature volumes and anchor-guided feature volumes, which are subsequently fed through (c) the pose estimation network. The proposed network initially predicts the 3D heatmaps from the keypoint feature volumes and then utilizes these 3D heatmaps along with the anchor-guided feature volumes to generate (d) the 3D keypoints for each person. If the 3D keypoints from the previous time step are available, they can be used to filter the 3D heatmaps. The entire network training does not require real image-3D keypoints pairs; instead, can be accomplished with only synthetic data.

on social motion, ignoring more challenging cases, such as sports and close interactions.

In recent years, several datasets containing close human interactions have been proposed. CHI3D [Fieraru et al. 2020] and ExPI [Guo et al. 2022], for instance, captured numerous two-person interaction actions with multi-view cameras. However, they lack subject diversity as they only include a few pairs of actors. The most recent interaction dataset Hi4D [Yin et al. 2023] registers neural implicit avatars to raw scans, enabling 4D tracking of both geometry and pose. However, all these datasets are restricted to indoor scenes.

Instead of capturing real data, some methods have attempted to apply synthetic data to overcome the issue of data deficiency. Some methods [Liu et al. 2022; Su et al. 2022] synthesize 2D heatmaps while most [Bazavan et al. 2021; Black et al. 2023; Mehta et al. 2018; Patel et al. 2021; Varol et al. 2017] synthesize RGB images. For instance, Surreal [Varol et al. 2017] utilizes MoCap poses [CMU Graphics Lab 2000] and textures [Robinette et al. 2002] to synthesize a single-person dataset. MuCO [Mehta et al. 2018] and AGORA [Patel et al. 2021] use synthetic methods to generate static multi-person datasets. To reduce sim-to-real gap, HSPACE [Bazavan et al. 2021] and BEDLAM [Black et al. 2023] synthesized large-scale photo-realistic dynamic video datasets with scan animation or physical simulation, but they only showed results on general scenes without close interactions.

Besides synthesizing data, some methods have applied view augmentation to improve generalizability. Drover et al. [2018] transformed 2D pose into a 3D pose and random projected it into a 2D pose in a new viewpoint. The artificially projected 2D pose was then evaluated by a discriminator using an adversarial approach. Chen et al. [2019] extended Drover et al. [2018] by lifting the projected 2D pose into a 3D pose, projecting it back to 2D with inverse transformation, and proposing a consistency loss for both 2D and 3D space. Wandt et al. [2021] explored view augmentation on multi-view data. However, these methods are not directly adaptable to multi-person scenarios.

### 3 TECHNICAL APPROACH

Our system takes as input the multi-view images of a scene with known camera parameters and outputs the 3D poses of all individuals in the scene, particularly for cases involving close interactions. To achieve this, we begin by extracting 2D keypoint heatmaps and estimating the 3D center of each individual (Section 3.1). Subsequently, we construct feature volumes for each person (Section 3.2) and feed them through our proposed network (Section 3.3), which is trained with only synthetic data (Section 3.4). The overview of our method is illustrated in Fig. 4.

#### 3.1 Center Estimation and Tracking

The center of each individual provides a rough estimate of their position within the scene. We use the pelvis point as the center of the body. From each image, 2D human keypoint heatmaps are extracted to identify potential 2D positions for the center points. Our goal is to select 2D candidates from different views and triangulate them to derive a valid 3D point. A valid 3D point should exhibit a minimal reprojection error when projected onto the selected views.

*Triangulation from 2D candidates.* Once we have these 2D candidates, they are arranged in descending order based on their scores. Beginning by selecting the two candidates with the highest scores (heatmap responses) from different views, we proceed to triangulate them to create a 3D point. We then project this point to all views and check the reprojection error. If the error is below the specified threshold, we consider this 3D point as valid. Otherwise, we proceed to select the next 2D candidate with the highest score. This iterative process, which involves the selection, triangulation, and subsequent evaluation of candidate positions, continues until all the 2D candidate positions have been evaluated.

*Tracking from previous frame.* For sequential input, we can streamline the procedure. We simply project the estimated centers of all individuals from the previous frame onto the current frame, retaining only those 2D points that meet the threshold condition.



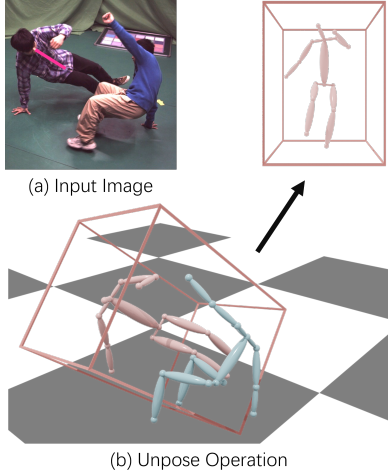


Fig. 5. **Coordinate transformation in 3D pose estimation.** We apply "unpose" operation to the estimated torso part (marked by the pink line in (a)), which is transformed into a standard space thus reducing the influence of global rotation.

Subsequently, selected 2D points are triangulated to derive the center coordinates for the current frame. These reconstructed centers are utilized to construct feature volumes for each individual. Our method is also applicable to the neck point since both the pelvis and neck points can be robustly detected and tracked. These two points serve as anchor points for subsequent stages.

### 3.2 Constructing Feature Volume

Previous methods usually discretize the 3D space into a fixed-size volume. They build a feature volume from 2D heatmaps or image features and employ a network to estimate the keypoint probabilities. In our approach, we strive for the network to be independent of actual image features. Hence, we opt to construct the feature volume using 2D heatmaps.

*Unposing transformation.* Before constructing the feature volume, we first "unpose" extreme poses encountered in real-world scenarios, such as lying down or rolling over, to a standard space (as illustrated in Fig. 5). The approach usually has minimal impact on common interaction datasets like CHI3D [Fieraru et al. 2020] and Hi4D [Yin et al. 2023], as most of the poses in these datasets involve standing. However, it is significantly beneficial when dealing with more complex actions. We use the estimated pelvis as the volume's center and the vector from the pelvis to the neck (as described in Sec. 3.1) as the reference vector. When constructing the volume, we rotate and translate the world coordinate system so that the transformed coordinate system is centered at the pelvis, with its z-axis aligned with the reference vector. The keypoint position in the standard coordinate system is then converted back to the original world coordinates via the inverse rotation and translation.

*Keypoint feature volume.* Given the unposed volume of each individual, we divide it into a discrete grid of size  $W \times H \times D$ . For each point  $\mathbf{x} \in \mathbb{R}^3$  within this volume, we project it onto the  $v$ -th

view using full-perspective projection and then bilinearly sample the heatmaps at its projected location. Subsequently, we combine the heatmap response vectors from different views into a global vector through averaging. The entire process can be formulated as:

$$\mathbf{F}_x = \frac{1}{V} \sum_v \mathbf{h}_v (\Pi_v(\mathbf{x}; \mathbf{K}_v, \mathbf{R}_v, \mathbf{t}_v)), \quad (1)$$

where  $\mathbf{K}_v$ ,  $\mathbf{R}_v$ , and  $\mathbf{t}_v$  are cameras intrinsics and extrinsics of  $v$ -th viewpoint and  $\Pi_v$  is the perspective projection function. Here,  $\mathbf{h}_v$  represents the mapping function that samples the response of each joint from the 2D heatmap.

*Anchor-guided feature volumes.* Simply building feature volumes around centers will introduce significant ambiguity when people closely interact with each other. This is because the feature volumes are nearly the same for each individual, as illustrated in Fig. 2. We address this ambiguity by employing the anchor points (pelvis and neck) since they provide positional information for the upper and lower body respectively. We utilize the anchor points of the  $i$ -th person, denoted as  $\mathbf{c}^i \in \mathbb{R}^{2 \times 3}$ , to extract the relevant features for that individual. Additionally, we use the anchor points of the other individuals, denoted as  $\mathbf{c}_o^i = \{\mathbf{c}^k \in \mathbb{R}^{2 \times 3} | k = 1, \dots, N, k \neq i\}$ , to suppress the responses of keypoints belonging to others.

We model the response of grid points to the anchor points using a Gaussian function. The positive response volume  $\mathbf{Z}^i \in \mathbb{R}^{2 \times W \times H \times D}$  is calculated as follows:

$$\mathbf{Z}^i = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{c}^i\|_2^2\right), \quad (2)$$

where  $\sigma$  represents the Gaussian radius. To account for the negative response, we calculate  $\mathbf{Z}^k \in \mathbb{R}^{2 \times W \times H \times D}$  for each element of  $\mathbf{c}_o^i$  using Eq. (2). This is followed by an element-wise maximum operation to obtain a fused field  $\mathbf{Z}_o^i \in \mathbb{R}^{2 \times W \times H \times D}$  computed as:

$$\mathbf{Z}_o^i = \max_k \mathbf{Z}^k, \quad (3)$$

where  $k$  denotes the index of other individuals. We call  $\mathbf{Z}^i$  and  $\mathbf{Z}_o^i$  *anchor-guided feature volumes*.

The keypoint feature volume and the anchor-guided feature volumes serve as inputs to the pose estimation network. Both can be synthesized using 3D human poses, which means that we can train our network without the need for actual images.

### 3.3 Pose Estimation Network

Previous approaches [Tu et al. 2020; Ye et al. 2022] typically feed a keypoint feature volume  $\mathbf{F}^i \in \mathbb{R}^{J \times W \times H \times D}$  of each person into the network to estimate the keypoint probability volume  $\mathbf{P}^i \in \mathbb{R}^{J \times W \times H \times D}$ . However, these methods struggle to accurately recover 3D poses in close interaction scenarios. This challenge often arises from the noise (multiple high responses for the same keypoint) and ambiguity (similar values for different people) present in the feature volume  $\mathbf{F}^i$ , as illustrated in Fig. 6.

*Two-stage network.* To address these issues, we propose a 3D *Heatmap Estimation Module* (HEM) and a *Keypoint Localization Module* (KLM) to enhance the network's understanding of the scene. In the first stage, we take  $\mathbf{F}^i$  as input and output a 3D heatmap volume  $\hat{\mathbf{H}}^i \in \mathbb{R}^{J \times W \times H \times D}$  for all appeared keypoints:

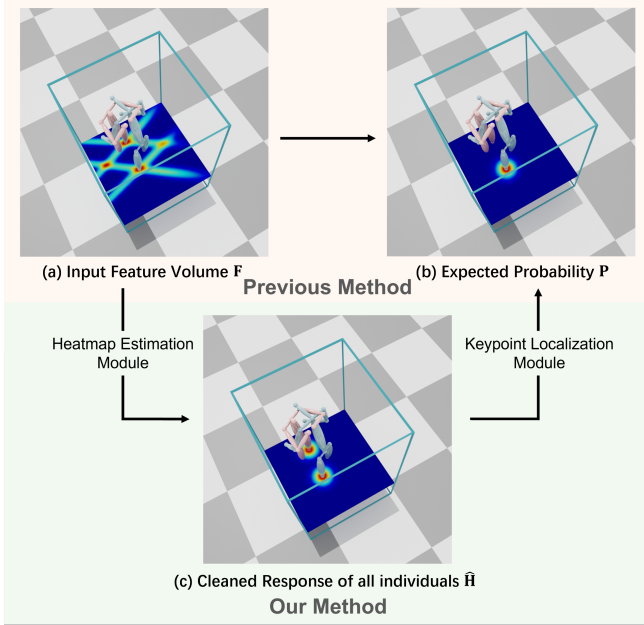


Fig. 6. **Two-stage design.** This image highlights the main difference between our approach and the previous methods in the field. Given the feature volume obtained through multiple viewpoints (a), the previous methods directly estimate the keypoint probability volume (b) of a target person. In contrast, we propose a two-stage method. The first Heatmap Estimation Module focuses on identifying and filtering out the noise present in the input feature volume and outputs a cleaned response volume of all individuals (c), while the second Keypoint Localization Module leverages the cleaned response volume and the conditional inputs to acquire the desired keypoint probability volume for each individual. This two-stage design allows the network to gain a better understanding of the scene.

$$\hat{\mathbf{H}}^i = \text{CNN}_{\text{HEM}}(\mathbf{F}^i). \quad (4)$$

This stage filters out erroneous responses and provides a well-regularized feature space for the subsequent stage. In the second stage, we concatenate  $\hat{\mathbf{H}}^i$  with anchor-guided feature volumes and feed them into 3D CNNs to robustly regress the keypoint probability volume  $\mathbf{P}^i$  for each person. The inclusion of the anchor-guided feature volumes helps the network learn to suppress keypoint responses from other individuals and focus on the target person.

The entire process can be formulated as:

$$\mathbf{P}^i = \text{CNN}_{\text{KLM}}(\hat{\mathbf{H}}^i, \mathbf{Z}^i, -\mathbf{Z}_o^i). \quad (5)$$

The final 3D coordinates can be calculated by taking the expectation of this volume as follows:

$$\hat{\mathbf{y}}_j^i = \sum_{l=1}^W \sum_{m=1}^H \sum_{n=1}^D \mathbf{P}_j^i(\mathbf{x}) \cdot \mathbf{x}. \quad (6)$$

**Temporal filtering.** For video input, we can further enhance the precision of the estimation by leveraging the estimated keypoints  $\{\hat{\mathbf{y}}_{j,t-1}^i | j = 0, \dots, J\}$  from the previous frame. Given the 3D heatmap volume estimated by HEM, we eliminate the incorrect responses

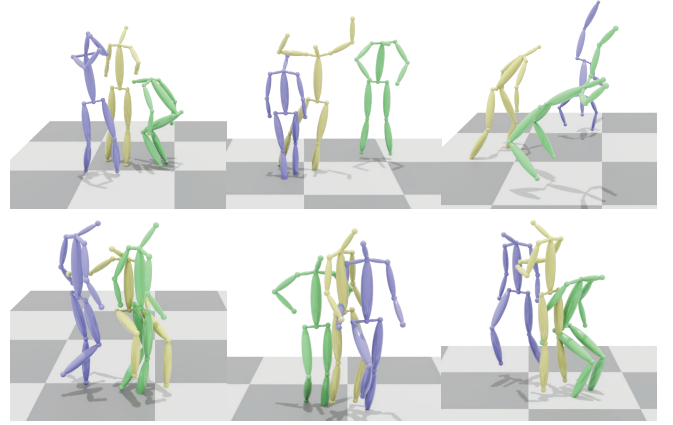


Fig. 7. **Synthesized multi-person 3D skeletons from the CMU MoCap [CMU Graphics Lab 2000] dataset.** The first row demonstrates some interactive actions sampled from the dataset, while the second row shows some synthesized close interactions.

within it by assuming that the movement distance of the  $j$ -th keypoint between two consecutive frames is less than a threshold  $r$ . This step can be formally described as:

$$\hat{\mathbf{H}}_j^i(\mathbf{x}) = \begin{cases} \hat{\mathbf{H}}_j^i(\mathbf{x}) & \text{if } \|\hat{\mathbf{y}}_{j,t-1}^i - \mathbf{x}\|_2 < r \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where  $\mathbf{x}$  is the grid point of 3D heatmap volume. We set  $r = 0.05\text{m}$  in our experiments.

The proposed pose estimation network can be trained end-to-end with 3D keypoint supervision. The temporal filtering module serves as a preprocessing step during inference.

### 3.4 Synthetic Training Data

While attention has been given to collecting interaction data [Fieraru et al. 2020; Guo et al. 2022; Yin et al. 2023], real-world multi-person motion datasets with diverse subjects, actions, and settings remain limited. This limitation arises from the difficulty of data collection, resulting in a significant gap in diversity and richness compared to other tasks. To alleviate the data scarcity issue, we propose a training strategy that relies only synthetic data.

Given the camera parameters of a test scene, we randomly sample  $N$  3D poses from publicly available MoCap datasets and project them onto each camera plane to obtain the corresponding 2D heatmaps. This process results in a set of 2D-3D pose pairs. We perform data augmentation at both the 3D and 2D levels to improve data diversity. Specifically, for each sampled 3D pose, we randomly rotate and place it within the scene. To synthesize close interactions, we manually move the placed poses towards each other (as shown in Fig. 7). At the 2D level, we simulate keypoint occlusion using two strategies: random viewpoint dropout and random 2D keypoint dropout. To simulate the uncertainty of actual 2D backbone inference (e.g., missing keypoints, false detections, imperfect Gaussian distribution, and noise), we perform augmentation on the generated heatmaps by randomly perturbing the values, positions, and shapes of Gaussians.

We obtain the 3D motion data from the CMU MoCap [CMU Graphics Lab 2000] dataset, which contains motion clips of 543.49 minutes. To reduce data redundancy, we select 671,790 poses based on the measure of movements. The selected poses cover a wide range of human actions. The impact of data synthesis strategies is analyzed by experiments in Section 5.7.

## 4 IMPLEMENTATION DETAILS

*Keypoint definition.* To be consistent with previous methods [Tu et al. 2020], we preserve the same definition of keypoints as used in earlier works, *i.e.*, the 15 keypoints defined in the CMU Panoptic dataset [Joo et al. 2015]. Other datasets used in our study are converted to follow this definition for evaluation purposes, by using the given SMPL [Loper et al. 2015] parameters and the SMPL model.

*2D keypoint heatmap estimator.* We train the HRNet [Wang et al. 2020b] on the COCO [Lin et al. 2014] dataset, requiring the network to output the keypoint responses of all individuals within a bounding box (consistent with the supervision of bottom-up methods). Off-the-shelf bottom-up 2D pose estimators (*e.g.*, OpenPose [Cao et al. 2017]) can also be applied in our approach.

*Pose estimation network.* Similar to previous works [Tu et al. 2020], we utilize a voxel-to-voxel 3D convolutional network as the basic module. For each individual, we construct a volume of  $2m \times 2m \times 2m$ , which is divided into a  $W \times H \times D$  grid where  $W = H = D = 32$ .

*Training.* We apply the Focal Loss [Lin et al. 2017] for heatmap estimation in the first stage, and apply the L1 Loss to the regressed keypoint coordinates  $\hat{y}_j^i$  in the second stage. The final loss function for the  $i$ -th person is computed as:

$$L^i = \lambda \text{FocalLoss}(\hat{H}^i, H^i) + \frac{1}{J} \sum_j |\hat{y}_j^i - y_j^i|, \quad (8)$$

where  $H^i$  is the ground-truth 3D keypoint heatmap of the  $i$ -th person, and  $y_j^i$  denotes the  $j$ -th ground truth keypoint coordinate of the  $i$ -th person. We use the Adam [Kingma and Ba 2015] optimizer with a learning rate of 0.0001 and apply exponential decay with a gamma value of 0.95. We train the model on synthetic data for 50 epochs with a batch size of 32. For hyper-parameters, we use  $\sigma = 0.05m$  in Eq. (2), and  $\lambda = 1$  in Eq. (8) for all experiments.

## 5 EXPERIMENTS

### 5.1 Datasets

We use public datasets AMASS [Mahmood et al. 2019], CHI3D [Fieraru et al. 2020], Hi4D [Yin et al. 2023] and CMU Panoptic [Joo et al. 2015] for training and evaluation.

AMASS [Mahmood et al. 2019] is a large and diverse database of human motion that includes multiple marker-based MoCap datasets. It uses MoSh++ to obtain the SMPL [Loper et al. 2015] parameters for each data point. We primarily use the largest CMU MoCap [CMU Graphics Lab 2000] part as our training data for all experiments. We use the SMPL model to convert the SMPL parameters into 3D keypoints. This dataset contains 96 subjects and 1983 motions. We remove frames with movements less than 0.05m and ultimately obtain 671,790 frames in total.

CHI3D [Fieraru et al. 2020] is an indoor dataset that focuses on close human interactions. It contains 631 multi-view sequences and 728,664 3D skeletons. Each sequence captures 2 people performing various actions from 4 different views with a resolution of  $900 \times 900$  at 50 fps. The dataset also provides the ground-truth (GT) SMPL-X [Pavlakos et al. 2019] parameters for one person (obtained from markers) and pseudo GT parameters for the other person (obtained from RGB images). We use ‘s02’ and ‘s04’ as the training sequences for the baseline methods and ‘s03’ as the test sequence for all methods since the actors in ‘s02’ and ‘s04’ are the same. We sample all sequences at 10fps and convert the provided SMPL-X parameters into 3D keypoints defined by COCO.

Hi4D [Yin et al. 2023] is the latest indoor close interaction dataset containing 100 sequences. Each sequence captures 2 actors from 8 different viewpoints with a resolution of  $1280 \times 940$  at 50 fps. The dataset provides the GT SMPL [Loper et al. 2015] parameters for each frame. We select ‘fight00’, ‘fight12’, ‘hug00’, ‘hug09’, ‘hug12’, ‘dance10’, ‘dance14’, ‘dance28’, ‘pose32’ and ‘pose37’ for evaluation. We use all frames within the selection and convert the SMPL parameters into COCO keypoints similar to CHI3D.

CMU Panoptic [Joo et al. 2015] is the largest real-world benchmark for multi-view multi-person 3D pose estimation. It contains 65 sequences and 1.5 million 3D skeletons with 30+ HD cameras. We use this dataset to evaluate our method on the setting of general multi-person poses. We follow Tu et al. [2020] testing our method on ‘160906\_pizza1’, ‘160422\_hagglng1’, ‘160906\_ian5’, and ‘160906\_band4’.

### 5.2 Baseline Methods

We compare our method against the latest multi-view multi-person 3D pose estimation methods from two categories: learning-based methods and association-based methods. The learning-based baselines include the voxel-based method VoxelPose [Tu et al. 2020] and its fast version Faster-VP [Ye et al. 2022], as well as a graph-based method Graph [Wu et al. 2021]. The association-based baselines include the top-down method MVPose [Dong et al. 2019] and the bottom-up method 4DA [Zhang et al. 2020].

### 5.3 Metrics

We utilize two metrics to evaluate the estimated 3D poses: 3D Percentage of Correct Keypoints (3DPCK) [%] and Mean Per Joint Position Error (MPJPE) [mm]. 3DPCK determines whether a keypoint is correctly estimated by measuring the Euclidean distance between the ground-truth (GT) position and the estimated joint position. If this distance is within a threshold, the estimation is considered as correct. Specifically, we match each estimated pose with the closest GT pose. If multiple estimates match with the same GT, only the one with the highest proposal score is viewed as a True Positive (TP), while the others are viewed as False Positives (FP). MPJPE calculates the Euclidean distance between the GT keypoint positions and the matched estimated keypoint positions.

### 5.4 Evaluation on Close-Interaction Datasets

The quantitative evaluation of our method and the baseline methods on the CHI3D [Fieraru et al. 2020] dataset is given in Tab. 1.



Table 1. **Evaluation on CHI3D [Fieraru et al. 2020]**. We report 3D Percentage of Correct Keypoints (3DPCK) with a threshold of 50mm here, so higher is better. Our approach achieves state-of-the-art results compared to the previous methods, surpassing even learning-based methods trained on the dataset by a large margin. ‘†’ indicates the methods that are trained on Panoptic [Joo et al. 2015] with 4 camera views close to CHI3D. ‘\*’ indicates the methods that are trained with synthetic data generated by their official code. ‘\*\*’ indicates the methods that are trained on the ‘s02’ and ‘s04’ sequences of CHI3D.

|                           | Method      | Grab         | Handshake | Hit          | HoldingHands | Hug          | Kick         | Posing       | Push         | All          |
|---------------------------|-------------|--------------|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Learning-based methods    | VoxelPose†  | 36.28        | 38.20     | 41.60        | 40.21        | 31.22        | 41.04        | 26.08        | 41.02        | 38.36        |
|                           | Faster-VP†  | 10.22        | 9.99      | 9.72         | 7.61         | 7.58         | 7.85         | 6.79         | 10.63        | 9.19         |
|                           | Graph†      | 28.70        | 28.89     | 30.97        | 26.14        | 20.40        | 32.62        | 18.00        | 30.31        | 28.33        |
|                           | VoxelPose*  | 82.78        | 86.55     | 84.93        | 89.01        | 64.14        | 84.47        | 72.99        | 86.82        | 82.40        |
|                           | Faster-VP*  | 81.91        | 85.57     | 84.66        | 91.99        | 65.03        | 84.52        | 71.59        | 85.37        | 82.22        |
|                           | VoxelPose** | 90.79        | 92.61     | 91.41        | 92.57        | 75.30        | 90.92        | 78.51        | 91.70        | 88.91        |
|                           | Faster-VP** | 89.93        | 92.40     | 90.35        | 96.05        | 67.42        | 89.07        | 77.48        | 90.70        | 87.45        |
| Graph**                   | 93.45       | <b>95.77</b> | 93.68     | 95.80        | 75.84        | 93.63        | 76.83        | 93.88        | 90.92        |              |
| Association-based methods | MVPose      | 74.42        | 77.65     | 75.48        | 84.48        | 57.86        | 75.01        | 69.33        | 76.55        | 74.16        |
|                           | 4DA         | 79.80        | 75.17     | 76.65        | 83.75        | 72.83        | 76.88        | 78.02        | 77.79        | 77.45        |
|                           | <b>Ours</b> | <b>95.68</b> | 95.28     | <b>94.92</b> | <b>97.40</b> | <b>88.14</b> | <b>94.30</b> | <b>92.16</b> | <b>95.23</b> | <b>94.30</b> |

*Comparison with learning-based methods.* Learning-based methods directly trained on the Panoptic [Joo et al. 2015] dataset exhibit poor performance on CHI3D. The results suggest that these methods struggle to generalize across different camera layouts. Therefore, we train VoxelPose [Tu et al. 2020] and Faster-VP [Ye et al. 2022] with synthetic heatmaps generated by their official code given the cameras of CHI3D, obtaining improved results. The best performance of learning-based methods is obtained by training using the real image-pose pairs from CHI3D. This implies that learning-based methods perform better with a sufficient amount of training data from the same scene. Compared to these methods, our method significantly outperforms them, demonstrating the effectiveness of our approach. Some qualitative comparisons are shown in Fig. 10.

*Comparison with association-based methods.* MVPose [Dong et al. 2019] relies on body-level matching and reconstruction, so it often struggles with close interactions. In contrast, 4DA [Zhang et al. 2020] initially reconstructs joints and then associates them in 3D, thus producing more reasonable skeletons, but its accuracy is still limited due to missing 2D detections. Compared with them, our method excels in reconstructing accurate poses in complex scenarios, as demonstrated in Fig. 11.

*Keypoint-level evaluations.* We evaluate our method on different keypoints separately to further investigate the accuracy of interacting body parts. Specifically, we evaluate 3DPCKs of keypoints that are more involved in human-human interactions (e.g., Shoulder, Elbow, Wrist, Hip, Knee, and Ankle) on a typical scenario ‘Hug’, as shown in Fig. 8. We find that the overall accuracy of lower-body keypoints (e.g., Hip, Knee, and Ankle) is better than upper-body keypoints (e.g., Shoulder, Elbow, and Wrist). This is because the hugging motion mostly involves upper-body parts.

*Generalization across datasets.* We further evaluate different methods on the Hi4D [Yin et al. 2023] dataset without training on this

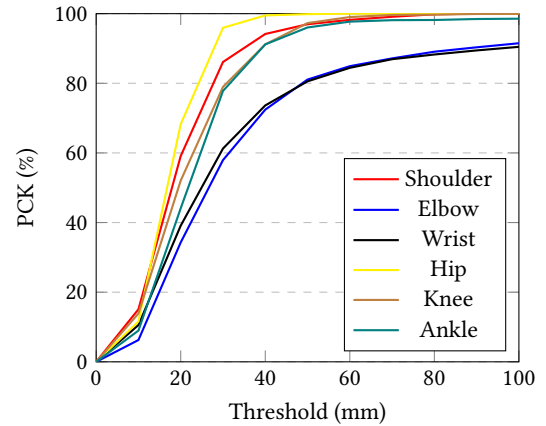


Fig. 8. **Keypoint-level evaluation on the ‘Hug’ of CHI3D [Fieraru et al. 2020]**. We report 3DPCKs with a threshold from 0 to 100mm on ‘Hug’. We can find that the 3DPCKs of ‘Elbow’ and ‘Wrist’ are larger than others. This is because the upper body experiences closer interaction than the lower body.

dataset. The quantitative evaluation is given in Tab. 2. Learning-based methods trained on CHI3D exhibit similar performances, slightly worse than association-based methods. This suggests that with more cameras (8 cameras in Hi4D versus 4 cameras in CHI3D), association-based methods can outperform learning-based ones. Despite this, our method continues to surpass both learning-based and association-based methods. Additionally, we present the 3DPCK curves with thresholds ranging from 0-100mm for all methods. As shown in Fig. 9, our method consistently outperforms others across all thresholds.

Table 2. **Evaluation on Hi4D [Yin et al. 2023]**. We report 3DPCKs with different thresholds (50mm, 100mm, and 200mm) and MPJPE. ‘§’ indicates the methods that are trained on CHI3D [Fieraru et al. 2020] and tested on Hi4D using the 4 views close to the training ones. ‘\*\*’ indicates the methods that are trained with synthetic data generated from CHI3D using their official code.

| Method                 | PCK@50↑      | PCK@100↑     | PCK@200↑     | MPJPE↓       |
|------------------------|--------------|--------------|--------------|--------------|
| VoxelPose <sup>§</sup> | 69.08        | 83.61        | 89.41        | 59.47        |
| Faster-VP <sup>§</sup> | 71.63        | 85.75        | 93.21        | 58.22        |
| Graph <sup>§</sup>     | 79.86        | 90.38        | 94.28        | 45.63        |
| VoxelPose*             | 83.56        | 89.96        | 92.28        | 42.00        |
| Faster-VP*             | 81.35        | 91.35        | 94.35        | 43.07        |
| 4DA                    | 87.83        | 98.24        | 99.48        | 31.60        |
| MVPose                 | 83.75        | 95.87        | 97.94        | 37.53        |
| <b>Ours</b>            | <b>98.29</b> | <b>99.55</b> | <b>99.68</b> | <b>20.28</b> |
| Ours(w/ $Y_{t-1}$ )    | 98.22        | 99.64        | 99.90        | 19.40        |

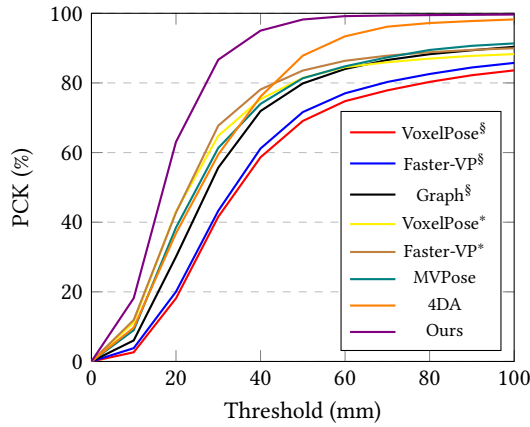


Fig. 9. **Evaluation on Hi4D [Yin et al. 2023] with tight thresholds**. We report 3DPCKs with a tight threshold from 0 to 100mm. The results show that our method outperforms others by a large margin even in tight thresholds. The notations of methods follow those in Tab. 2.

### 5.5 Evaluation on General Multi-Person Tracking Dataset

To demonstrate the generality of our approach, we evaluate it on the Panoptic [Joo et al. 2015] dataset which captures more regular social activities in a studio and serves as a common benchmark for previous research [Tu et al. 2020; Wu et al. 2021; Ye et al. 2022]. The results are given in Tab. 3, where the numbers of the previous methods are adopted from their papers [Tu et al. 2020; Wang et al. 2021; Wu et al. 2021; Ye et al. 2022] using their original data partition and evaluation protocols. As this dataset contains very few examples of close interaction, previous methods also perform well on it. Our method achieves comparable performance even without the use of real image-3D pose pairs during training. Graph [Wu et al. 2021] and MvP [Wang et al. 2021] exhibit better results in the  $AP_{25}$  metric. This improvement may be attributed to their use of multi-layered image features extracted by the 2D CNN. If we train our network using the

Table 3. **Evaluation on Panoptic [Joo et al. 2015]**. We report Average Precision (AP) with thresholds of 25, 50, and 100mm, where the higher values indicate better performance. Ours is trained using synthetic data, while Ours\* is trained using heatmaps generated from images.

|                             | Method    | $AP_{25}$ ↑  | $AP_{50}$ ↑  | $AP_{100}$ ↑ |
|-----------------------------|-----------|--------------|--------------|--------------|
| Heatmap-based methods       | VoxelPose | 83.59        | 98.33        | <b>99.76</b> |
|                             | Faster-VP | 85.22        | 98.08        | 99.32        |
|                             | Ours      | 86.16        | <b>98.70</b> | 99.49        |
|                             | Ours*     | <b>90.10</b> | 98.57        | 99.34        |
| Image-feature-based methods | Graph     | <b>94.00</b> | <b>98.93</b> | <b>99.76</b> |
|                             | MvP       | 92.28        | 96.60        | 97.45        |

keypoint heatmaps extracted from the real images, our  $AP_{25}$  exhibits an improvement (Ours\* in the table), where there is a marginal change in  $AP_{50}$  and  $AP_{100}$ . This suggests that incorporating real heatmaps for training can enhance the network’s precision.

### 5.6 Ablation Study

*Core components.* We conduct ablation studies to evaluate the impact of our core components, *i.e.*, *3D heatmap supervision* and *conditional inputs*. The experiments are evaluated on the CHI3D [Fieraru et al. 2020] dataset. We use ground-truth centers here to mitigate the influence of center estimation. The results are shown in Tab. 4. For *3D heatmap supervision*, we train an ablated model without supervision of the 3D heatmap volume, *i.e.*, the network is not explicitly required to output 3D heatmaps of all individuals. We find that the removal of heatmap supervision decreases the performance in all metrics, indicating the importance of heatmap supervision. As for *conditional inputs*, we train an ablated model where conditional inputs, *i.e.*, the anchor-guided feature volumes, in the keypoint localization module are eliminated. The results show that the removal of conditional inputs leads to degraded performance, emphasizing the role of the conditional inputs in solving the ambiguity arising from the close interaction.

Table 4. **Ablation study on CHI3D [Fieraru et al. 2020]**. We report the MPJPE, PCK@50 for all scenarios, and PCK@50 for the ‘Hug’ scenario. This table highlights the significance of 3D heatmap supervision and conditional inputs and also shows the superior performance achieved with ground-truth 2D heatmaps.

|                            | MPJPE↓ | PCK↑  | PCK@50(Hug)↑ |
|----------------------------|--------|-------|--------------|
| w/o 3D Heatmap Supervision | 24.53  | 91.58 | 85.05        |
| w/o Conditional Inputs     | 22.02  | 95.56 | 90.42        |
| w GT 2D Heatmap            | 8.77   | 99.92 | 99.46        |
| Ours                       | 18.78  | 97.12 | 93.28        |

*Heatmap used during inference.* We explore the performance difference between using GT heatmaps and estimated heatmaps during the inference phase in Tab. 4. Experimental results show a noticeable improvement when using GT heatmaps during inference, indicating that the heatmap estimation of our pre-trained model still has room



Fig. 10. **Qualitative comparison with learning-based methods on CHI3D [Fieraru et al. 2020] and Hi4D [Yin et al. 2023].** Our method exhibits superior accuracy in reconstructing 3D poses without using any real paired training data compared to other learning-based methods, particularly in challenging scenarios. The red box indicates the erroneous keypoints estimated by learning-based methods and even existed in the pseudo ground truth in CHI3D.



for improvement. Enhancements to the 2D heatmap estimation can effectively boost the overall performance of our approach.

*Number of views.* We evaluate the impact of the number of views on the Hi4D [Yin et al. 2023] dataset, as shown in Tab. 5. As the number of viewpoints decreases, the error of the estimated center and the overall MPJPE increase. Using the GT center estimation as an input reduces the error to a certain extent. This suggests that more accurate center estimation results in better overall performance.

Table 5. **Comparison of pose estimation performance with different numbers of views on Hi4D [Yin et al. 2023].** This table compares the performance of our proposed 3D pose estimation method with different numbers of views (4, 6, and 8). The performance metrics are MPJPE and MPJPE with ground-truth root (MPJPE w/ GT root). The lower values indicate better performance.

|         | Center Error↓ | MPJPE↓ | MPJPE w/ GT root↓ |
|---------|---------------|--------|-------------------|
| 4 views | 29.78         | 28.85  | 24.30             |
| 6 views | 23.01         | 21.50  | 16.07             |
| 8 views | 20.16         | 19.22  | 13.79             |

### 5.7 Evaluation of Synthetic Training Strategies

We conduct experiments on the CHI3D [Fieraru et al. 2020] dataset to evaluate the influence of our synthetic training strategies. The results are shown in Tab. 6.

*Quantity of synthetic data.* To investigate the impact of synthetic data size on model performance, we train our model using only half or a quarter of the CMU MoCap [CMU Graphics Lab 2000] dataset. Additionally, we test the performance of using the combined CMU MoCap and the BMLMoVi [Ghorbani et al. 2020] data from the AMASS [Mahmood et al. 2019] dataset for training. The experiments show that using less training data leads to performance decline, but the overall performance is still close to that of using the full data. Adding the extra BMLMoVi data does not bring a significant performance improvement, which suggests that the CMU MoCap dataset provides a sufficient variety of motion data.

*Different number of subjects in the scene.* We sample two subjects from CHI3D during training. If we sample only one subject per frame, there is a significant performance drop. If we sample more subjects each time (e.g., five subjects), the performance is close to that of two subjects.

*Data augmentation.* We further evaluate the effectiveness of 2D and 3D data augmentation. Without heatmap augmentation, the performance on the test set decreases significantly. If the center augmentation is not used, the performance degrades slightly.

### 5.8 Applicability to Large-Scale Scenes

To verify the generalizability of our method to large-scale scenes, we evaluate it on a dataset from a basketball court. This dataset was collected using 28 calibrated RGB cameras that fully encircle the basketball court. The data represents a typical basketball game with 10 players.

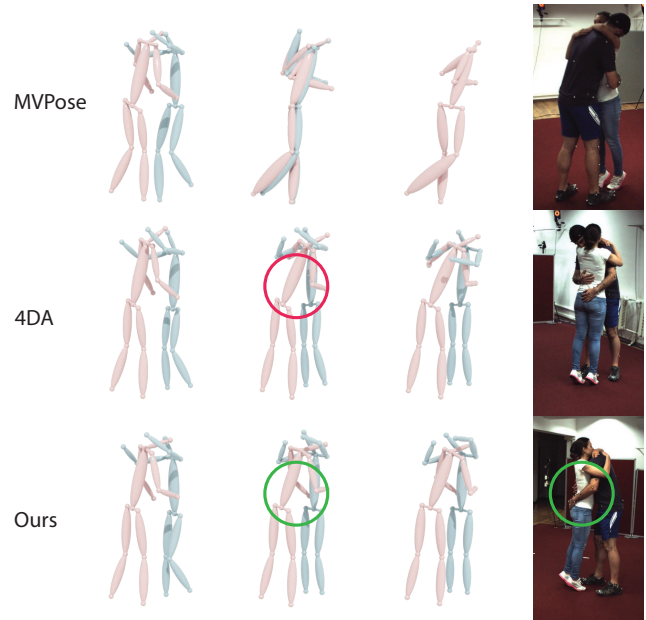


Fig. 11. **Qualitative comparison with association-based methods.** This figure compares the efficacy of our method and two pose estimation methods at varying interaction distances. We render skeletons using the viewpoint of the image in the first row. The traditional top-down method MV-Pose performs well at long distances but fails at close range. The bottom-up method 4DA excels at close range, though it still fails to reconstruct some keypoints as shown in the red circle. In contrast, our method accurately reconstructs poses in this complex scenario, outperforming the other two methods.

Table 6. **Comparison between multiple synthetic training strategies on CHI3D [Fieraru et al. 2020].** This table demonstrates that the CMU MoCap [CMU Graphics Lab 2000] dataset provides sufficient diversity for close-interaction pose estimation. Both multi-person sampling during synthetic data generation and heatmap augmentation are beneficial for our method.

|                               | MPJPE↓       | PCK↑         |
|-------------------------------|--------------|--------------|
| CMU MoCap + MoVi              | 18.92        | <b>97.17</b> |
| 1/2 CMU MoCap                 | 18.96        | 97.03        |
| 1/4 CMU MoCap                 | 19.50        | 96.87        |
| # of subjects = 1             | 23.97        | 93.73        |
| # of subjects = 5             | 18.80        | 97.13        |
| w/o 2D heatmap augmentation   | 51.99        | 76.81        |
| w/o center augmentation       | 18.87        | 97.04        |
| CMU MoCap + # of subjects = 2 | <b>18.78</b> | 97.12        |

Our method can be easily applied to this scenario, using the known camera parameters to synthesize training data. To cope with the issue of 2D human estimation in large scenes during testing, we use an off-the-shelf object detector [Jocher 2020] to obtain 2D human centers in the entire image, from which the 3D human centers

are reconstructed. Then, we project the 3D bounding box of each person back to the image, followed by cropping the original image. The cropped images are then used to estimate the 2D heatmaps, which serve as input for the our pose estimation network. The input images and reconstruction results are shown in Fig. 12. For such a large-scale scene, our method is also able to handle close-range interactions between people effectively.

## 6 APPLICATION IN NOVEL VIEW SYNTHESIS

Our method furnishes an advantageous starting point for many downstream tasks. Recently, novel view synthesis of dynamic humans from sparse views has been explored in [Liu et al. 2021; Mihajlovic et al. 2022; Peng et al. 2021; Shuai et al. 2022; Weng et al. 2022]. These methods adopted the SMPL [Loper et al. 2015] model or keypoints as the geometric prior to associate appearance information across different frames and learned a NeRF [Mildenhall et al. 2021] for rendering. We use the recent SMPL-based method [Shuai et al. 2022] for demonstration. The 3D skeletons recovered by our method are directly used to fit SMPL models. Subsequently, we employ multi-view images along with the SMPL parameters to train a dynamic radiance field following [Shuai et al. 2022], thereby facilitating the rendering of novel views, instance masks and depth maps of the scene, as demonstrated in Fig. 13. Contrasting with the data acquisition technique of the Hi4D dataset [Yin et al. 2023], our method requires merely eight RGB cameras, obviating the necessity for a high-cost scanning system. This simplifies the data acquisition process significantly while producing high-quality 3D reconstruction.

## 7 LIMITATION AND FUTURE WORK

While our method has shown promising results, there are some limitations that open up avenues for further improvements. First, our method currently only takes 2D keypoint heatmaps as input to directly output 3D keypoint coordinates. This could be enhanced by incorporating more 2D features into the input, such as the Part Affinity Field (PAF) from OpenPose [Cao et al. 2017]. This additional feature provides limb link information, thus potentially improving pose estimation, particularly in complex multi-person scenarios. Another limitation lies in the expressiveness of our output. Our method currently only outputs body keypoints, which may not fully capture the intricacy of human motion. As such, future work could consider fitting human body models from the estimated 3D keypoints with surface or contact losses and incorporating additional hand and facial keypoints. Lastly, our method does not involve motion prior during the training phase. We only utilize temporal input during the inference stage. Future work could incorporate temporal motion prior learning [Rempe et al. 2021] and spatial motion prior learning [Adeli et al. 2020; Guo et al. 2022; Katircioglu et al. 2021]. Such enhancement could potentially allow the model to gain more understanding of motion patterns and improve the overall accuracy of the pose estimation.

## 8 CONCLUSION

This paper introduces a novel system for multi-person 3D pose estimation from multi-view images, with a specific focus on scenarios

involving close interactions. Our technical contribution lies in the conditional 3D volumetric network along with its corresponding training and inference strategies. We demonstrated through rigorous experiments that our method significantly outperforms previous approaches, exhibiting robustness and generalization across a variety of scenarios. We expect this work will facilitate more future research in related areas and real applications that require motion capture or dynamic reconstruction of closely interacting humans.

## ACKNOWLEDGMENTS

The authors would like to acknowledge support from NSFC (No. 62172364) and the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

## REFERENCES

- Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezaatfoghi. 2020. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6033–6040.
- Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2021. HSPACE: Synthetic parametric humans animated in complex environments. *arXiv preprint arXiv:2112.12867* (2021).
- Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 2014. 3D pictorial structures for multiple human pose estimation. In *CVPR*. 1669–1676.
- Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 2015. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 10 (2015), 1929–1942.
- Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. 2020. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *CVPR*. 6856–6865.
- Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 2023. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *CVPR*. 8726–8737.
- Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2022. HuMMAN: Multi-modal 4d human dataset for versatile sensing and modeling. In *ECCV*. Springer, 557–577.
- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*. 7291–7299.
- Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. 2022. Multi-Person 3D Pose and Shape Estimation via Inverse Kinematics and Refinement. In *ECCV*. Springer, 660–677.
- Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, Nikolaos Zioulis, Shishir Subramanyam, Bart Kevelham, Caecilia Charbonnier, Pablo Cesar, Dimitrios Zarpalas, et al. 2020. HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media. *IEEE Access* 8 (2020), 176241–176262.
- Ching-Hang Chen, Amrith Tyagi, Amit Agrawal, Dylan Drover, Rohith Mv, Stefan Stojanov, and James M Rehg. 2019. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*. 5714–5724.
- CMU Graphics Lab. 2000. CMU Graphics Lab Motion Capture Database. <http://mocap.cs.cmu.edu/>.
- Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. 2019. Fast and robust multi-person 3d pose estimation from multiple views. In *CVPR*. 7792–7801.
- Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Amrith Tyagi, and Cong Phuoc Huynh. 2018. Can 3d pose be learned from 2d projections alone?. In *ECCVW*. 78–94.
- Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. 2020. Three-dimensional reconstruction of human interactions. In *CVPR*. 7214–7223.
- Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. 2021a. Learning complex 3d human self-contact. In *AAAI*. 1343–1351.
- Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. 2021b. AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training. In *CVPR*. 9919–9928.
- Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. 2021c. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *NeurIPS* 34 (2021), 19385–19397.
- Saeed Ghorbani, Kimia Mahdavian, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. 2020. MoVi: A Large Multipurpose Motion and Video Dataset. *Borealis*. <https://doi.org/10.5683/SP2/JRHDRN>

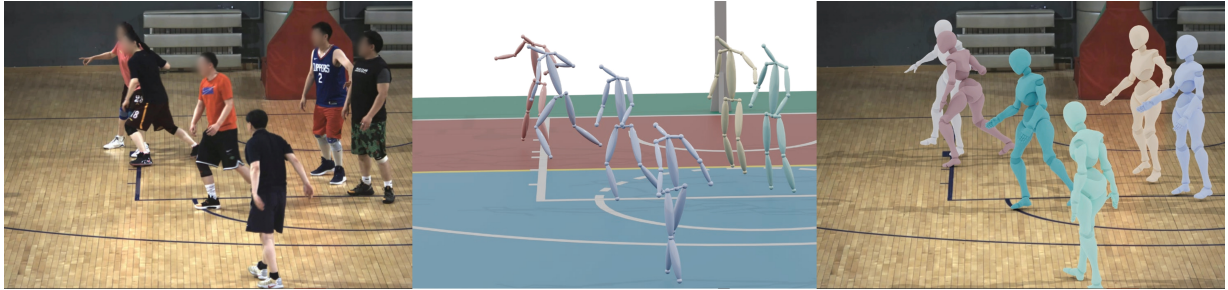


Fig. 12. This figure presents the results of our pose estimation and character animation on a large basketball court. Compared to indoor data, the basketball data has a larger scale and more people. Our method can perform pose estimation without the need for labeled training data on the basketball court. The results of pose estimation can be used for downstream applications such as motion analysis, character animation, and augmented reality.



Fig. 13. This figure showcases the results of free-viewpoint rendering achieved by learning a neural radiance field [Shuai et al. 2022] on the Hi4D [Yin et al. 2023] dataset using eight viewpoints of RGB images and our estimated 3D poses. The experimental results demonstrate that our method can effectively assist downstream tasks, including novel view synthesis (top), instance segmentation (middle), and depth estimation (bottom).

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

Congzhenhao Huang, Shuai Jiang, Yang Li, Ziyue Zhang, Jason Traish, Chen Deng, Sam Ferguson, and Richard Yi Da Xu. 2020. End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. In *ECCV*. Springer, 477–493.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1325–1339.

Karim Isakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. Learnable triangulation of human pose. In *ICCV*. 7718–7727.

Glenn Jocher. 2020. *Ultralytics YOLOv5*. <https://doi.org/10.5281/zenodo.3908559>

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*. 3334–3342.

Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. 2021. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*. IEEE, 42–52.

Isinsu Katircioglu, Costa Georgantas, Mathieu Salzmann, and Pascal Fua. 2021. Dyadic human motion prediction. *arXiv preprint arXiv:2112.00396* (2021).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *CVPR*. 2252–2261.

Jiahao Lin and Gim Hee Lee. 2020. Hdnet: Human depth estimation for multi-person camera-space localization. In *ECCV*. Springer, 633–648.

Jiahao Lin and Gim Hee Lee. 2021. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *CVPR*. 11886–11895.



- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*. 2980–2988.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.
- Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics* 40, 6 (2021), 16 pages.
- Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. 2022. Explicit Occlusion Reasoning for Multi-person 3D Human Pose Estimation. In *ECCV*. Springer, 497–517.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics* 34, 6 (Nov 2015), 16 pages.
- Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*. 5442–5451.
- Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. In *3DV*. IEEE. [http://gvv.mpi-inf.mpg.de/3dhp\\_dataset](http://gvv.mpi-inf.mpg.de/3dhp_dataset)
- Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Transactions on Graphics* 39, 4 (July 2020), 17 pages.
- Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*. 120–130.
- Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *ECCV*. Springer, 179–197.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. 2019. Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image. In *ICCV*. 10133–10142.
- Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *WACV*. 53–60.
- Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. 2021. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*. 13468–13478.
- Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*. 10975–10985.
- Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*. 9054–9063.
- Zhongwei Qiu, Yang Qiansheng, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. 2023. PSVT: End-to-End Multi-person 3D Pose and Shape Estimation with Progressive Video Transformers. In *CVPR*.
- Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. 2021. Humor: 3d human motion model for robust pose estimation. In *ICCV*. 11488–11499.
- Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoferlin, and Dennis Burnsides. 2002. Civilian American and European surface anthropometry resource (CAESAR), final report, volume I: Summary. *Sytronics Inc Dayton Oh* (2002).
- Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. 2022. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH*. 1–10.
- Leonid Sigal, Alexandru O Balan, and Michael J Black. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision* 87, 1-2 (2010), 4.
- Jiajun Su, Chunyu Wang, Xiaoxuan Ma, Wenjun Zeng, and Yizhou Wang. 2022. VirtualPose: Learning Generalizable 3D Human Pose Models from Virtual Data. In *ECCV*. Springer, 55–71.
- Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*. 13243–13252.
- Matt Trumble, Andrew Gilbert, Charles Malleon, Adrian Hilton, and John Collomosse. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *BMVC*.
- Hanyue Tu, Chunyu Wang, and Wenjun Zeng. 2020. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *ECCV*. Springer, 197–212.
- Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from synthetic humans. In *CVPR*. 109–117.
- Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*. 601–617.
- Bastian Wandt, Marco Rudolph, Petriša Zell, Helge Rhodin, and Bodo Rosenhahn. 2021. CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild. In *CVPR*.
- Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. 2020a. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*. Springer, 242–259.
- Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020b. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2020), 3349–3364.
- Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. 2021. Direct Multi-view Multi-person 3D Human Pose Estimation. *NeurIPS* 34 (2021), 13153–13164.
- Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, and Si Liu. 2022. Distribution-aware single-stage models for multi-person 3D pose estimation. In *CVPR*. 13096–13105.
- Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*. 16210–16220.
- Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. 2021. Graph-based 3d multi-person pose estimation using multi-view images. In *ICCV*. 11148–11157.
- Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *CVPR*. 6184–6193.
- Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. 2022. Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection. In *ECCV*. Springer, 142–159.
- Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2023. Decoupling Human and Camera Motion from Videos in the Wild. In *CVPR*.
- Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. 2023. Hi4D: 4D Instance Segmentation of Close Human Interaction. In *CVPR*. 17016–17027.
- Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Soo Park. 2021. Humbi: A large multiview dataset of human body expressions and benchmark challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2021), 623–640.
- Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. 2022. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*. 11038–11049.
- Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 2020. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*. 1324–1333.
- Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. 2020. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*. Springer, 550–566.
- Zhize Zhou, Qing Shuai, Yize Wang, Qi Fang, Xiaopeng Ji, Fashuai Li, Hujun Bao, and Xiaowei Zhou. 2022. QuickPose: Real-time Multi-view Multi-person Pose Estimation in Crowded Scenes. In *SIGGRAPH*. 1–9.