ContactField: Implicit Field Representation for Multi-Person Interaction Geometry

Hansol Lee
hansollee@kist.re.kr

Tackgeun You
tackgeun.you@kist.re.kr

Hansoo Park
hansupark@kist.re.kr

Woohyeon Shim
20211362@sungshin.ac.kr

Sanghyeon Kim
sangkim98@hanwha.com

Hwasup Lim
hslim@kist.re.kr

Center for Artificial Intelligence Research
Korea Institute of Science and Technology
Seoul, South Korea

Abstract

We introduce a novel implicit field representation tailored for multi-person interaction geometry in 3D spaces, capable of simultaneously reconstructing occupancy, instance identification (ID) tags, and contact fields. Volumetric representation of interacting human bodies presents significant challenges, including inaccurately captured geometries, varying degrees of occlusion, and data scarcity. Existing multi-view methods, which either reconstruct each subject in isolation or merge nearby 3D surfaces into a single unified mesh, often fail to capture the intricate geometry between interacting bodies and exploit on datasets with many views and a small group of people for training. Our approach utilizes an implicit representation for interaction geometry contextualized by a multi-view local-global feature module. This module adeptly aggregates both local and global information from individual views and interacting groups, enabling precise modeling of close physical interactions through dense point retrieval in small areas, supported by the implicit fields. Furthermore, we develop a synthetic dataset encompassing diverse multi-person interaction scenarios to enhance the robustness of our geometry estimation. The experimental results demonstrate the superiority of our method to accurately reconstruct human geometries and ID tags within three-dimensional spaces, outperforming conventional multi-view techniques. Notably, our method facilitates unsupervised estimation of contact points without the need for specific training data on contact supervision.

1 Introduction

Accurate 3D representations of multi-person interactions have critical applications in virtual reality, augmented reality, robotics, and surveillance, as human subjects are central to a variety of content and tasks. In particular, modeling interactions involving multiple individuals in close proximity has gathered attention as the modeling of individual humans and simple group activities has matured. However, the precise estimation and reconstruction of 3D human body poses and shapes in close interaction scenarios present significant challenges, mainly due to occlusion, which complicates accurate reconstruction.

The Skinned Multi-Person Linear (SMPL) model [2], one of the most well-known explicit models, has been extensively utilized not only for individual human models [37, 20, 6] but also in multi-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

person scenarios [8, 4, 39, 41, 40]. However, as an unclothed human model, it struggles to depict local details such as clothing and hairstyles. Additionally, methods using the SMPL model need separate parameter optimization for each person in a scene, which requires intricate coordination when modeling close interactions between individuals. Implicit representations [28, 21, 22] have been researched as alternatives to the SMPL model. Implicit models, with their higher degrees of freedom, are better suited for clearly expressing local details. Successful modeling of various scenes involving individual human avatars and their detailed local features in multi-person scenarios has been achieved in [43, 3, 23]. Nevertheless, the high degrees of freedom inherent to implicit models demand sophisticated neural architecture designs capable of handling multi-view image features and require high-quality data for training.

To address these challenges, we introduce a novel approach that represents multi-person interaction geometries by simultaneously dealing with geometry, identity, and contact fields in scenes without the need for a prior explicit model such as SMPL. Our proposed implicit field is optimized to estimate both the occupancy and identification (ID) fields, distinguishing each person in 3D space and modeling the interaction geometry. This approach enables the consideration of complex interactions between individuals while preserving the spatial information of each individual.

Furthermore, we utilize a multi-view feature transformer [5] and a global scene feature extraction transformer [34, 35, 36] to construct a 3D scene representation, addressing one of the biggest challenges in reconstructing close interactions: dealing with occlusions. By taking into account the global scene features for each point and leveraging latent 3D scene representations and Transformer architecture, we enhance the ability to infer information about occluded parts, which cannot be achieved with standard multi-view images alone. The Transformer, utilizing context provided by positional encoding and 3D scene representations, can infer the structure and position of occluded parts by learning from visible parts and their spatial relations. This capability is crucial for accurately reconstructing models of each person in the scene, even when direct visual information is lacking.

Additionally, we develop a synthetic dataset for multi-person interaction that includes interactions among 2, 3, and 4 individuals, which helps to address more diversity of characters and complex group dynamics in the scene.

Our experiments validate the superiority of our approach over existing methods, demonstrating its capability to accurately reconstruct and assign tag values in 3D space. Our contributions can be summarized as follows:

- We introduce a novel implicit field representation for multiple people in close interaction scenarios that simultaneously estimates multi-person geometries as occupancy fields, ID fields, and contact fields, thereby preserving their spatial relationships and capturing their interactions.
- Our method employs a novel multi-view local-global feature module coupled with a global scene features extraction technique, leveraging latent 3D scene representations to reconstruct individual geometries and assign ID values and contacts in complex, occluded scenarios.
- We demonstrate that our method can reconstruct 3D multi-person figures more effectively than existing methods. Also, we have created a synthetic dataset that models interactions among 2 to 4 individuals to enhance the understanding of group dynamics in close interactions.

The rest of the paper is organized as follows. We review the related works on 3D human representation in Section 2. We explain our method in Section 3 and demonstrate the effectiveness of the method on two datasets in Section 4. We conclude the paper in Section 5.

2 Related Work

Reconstructing 3D human models from RGB images or creating human avatars has been a longstanding challenge. An explicit model, the Skinned Multi-Person Linear model (SMPL) [18], dominates the human avatar research by serving as a canonical 3D human model [37, 15, 14, 30, 20, 6]. However, SMPL is an unclothed human model and is limited in its ability to capture local details such as clothing and hairstyles. Hence, implicit representations, including signed distance fields (SDF) [28] and occupancy fields [44], have gathered attention from the community. Pixel-aligned

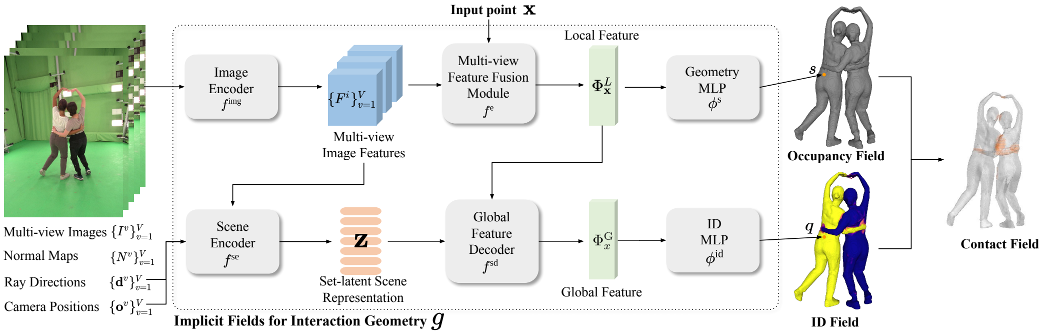

Input point $\mathbf{x}$
Local Feature
Image Encoder f^{img}
$\{F^i\}_{v=1}^V$
Multi-view Image Features
Multi-view Feature Fusion Module f^{e}
$\Phi_{\mathbf{x}}^L$
Geometry MLP ϕ^{s}
s
Occupancy Field
Multi-view Images $\{I^v\}_{v=1}^V$
Normal Maps $\{N^v\}_{v=1}^V$
Ray Directions $\{\mathbf{d}^v\}_{v=1}^V$
Camera Positions $\{\mathbf{o}^v\}_{v=1}^V$
Scene Encoder f^{se}
$\mathbf{Z}$
Set-latent Scene Representation
Global Feature Decoder f^{sd}
Φ_x^{G}
Global Feature
ID MLP ϕ^{id}
q
ID Field
Contact Field
Implicit Fields for Interaction Geometry g

Figure 1: The overall framework of our method. We compute the local and global features from a set of multi-view images and its camera parameters through the proposed multi-view feature local-global transformer. We use local feature to estimate occupancy and global feature to estimate ID at a given point $\mathbf{x}$. From the occupancy and ID fields, we estimate the contact field, as detailed in Section 3.4

feature encoders [32, 33, 13, 1, 16] contextualize the implicit fields by projecting image features onto 3D coordinates using the camera parameters to enhance 3D human avatar reconstruction. Utilizing multi-view RGB images effectively mitigates occluded parts of human in single-view RGB images. DoubleField [38] integrates neural surface and multi-view based radiance fields to represent 3D geometry and the appearance of humans.

Multi-person 3D reconstruction presents unique challenges not encountered in single-person scenarios, such as occlusion, where subjects may obstruct each other's visibility. This task requires not only a detailed understanding of spatial relationships among individuals but also the preservation of individuality while representing their interactions. This process needs precise reconstruction and a deep interpretation of interactions. MPSD [23] employs an implicit approach for the 3D reconstruction of each person, utilizing 6-DOF spatial position estimation within the global scene space. Their method enables multi-person 3D reconstruction from a single image, effectively tracking the locations of individuals. However, because this method primarily addresses data where individuals are spaced widely apart, it does not adequately handle scenarios involving people in close interaction.

Close interaction scenarios [43, 42, 3] are a crucial challenge in multi-person representation. DeepMultiCap [43] incorporates an attention module and temporal fusion to produce high-fidelity 3D models, relying on the SMPL model for prior segmentation, prediction, and location determination. However, this approach faces challenges, as the accuracy of 3D reconstruction and spatial estimation is bounded by the performance of an initial SMPL prediction, and it requires multiple optimization steps for each individual instance. The work [3] leverages a single-view image and gDNA [7] generative model to generate 3D geometry, refining the spatial positioning of each individual based on contact information. Considering that both methods sequentially reconstruct 3D geometry for each individual, accurately representing multiple subjects in close interaction remains a challenge. Also, existing methods for 3D reconstruction of close human interactions often suffer from data scarcity. The introduction of Hi4D [42] marks a leap forward, offering detailed 4D textures and essential data for studying two-person interactions. However, Hi4D [42] includes only two people, while the dataset described by MultiHuman [43] features up to three. In these two datasets, individuals are captured either separately or together, often overlapping or passing through each other.

Our method innovatively addresses these complexities by employing a transformer-based architecture that integrates multi-view feature fusion with global scene representation, allowing the simultaneous and dynamic reconstruction of multiple interacting individuals. This approach not only captures the detailed geometries of each person but also maintains their unique identities and spatial relationships, even in scenarios with significant occlusions and close physical interactions. Additionally, we created a synthetic dataset that models interactions among 2, 3, and 4 individuals to address even more complex group dynamics.

3 Method

Our method introduces a novel representation of multi-person interaction geometry by combining 3D reconstruction with identification of multiple individuals in close interaction from multi-view images. We achieve this by estimating two key fields: occupancy and ID. The occupancy estimation is crucial for reconstructing the geometry of each person, while the ID estimation facilitates the identification of individuals within the 3D space. We leverage transformer architectures to tackle the significant challenge of occlusions common in close interaction scenarios. Our architecture first extracts local features and then computes global features by integrating these local features with 3D scene representations. This approach ensures robust integration of local and global information across multiple perspectives, enabling a comprehensive understanding and reconstruction of the complex configurations of scenes. Figure 1 represents our process pipeline, encapsulating the essence of our approach.

3.1 Implicit Fields for Multi-Person Interaction Geometry

Our method utilizes implicit functions to represent interaction geometry in 3D space. Implicit functions uniquely define the surface of a 3D object by specifying a level set within a field [21]. This representation allows for a continuous definition of the surface, enabling the precise reconstruction of complex geometries. Alongside the geometric reconstruction, we introduce a novel approach to model the ID and contact fields within the same 3D space, providing a method for distinguishing individual entities in closely interacting scenarios.

Our model takes a query point $\mathbf{x} \in \mathbb{R}^3$ to predict two key attributes: the occupancy s, and the identification q for each query point. We extract multi-view local and global features, denoted as $\Phi_{\mathbf{x}}(\cdot) = \left[\Phi_{\mathbf{x}}^{\mathrm{L}}, \Phi_{\mathbf{x}}^{\mathrm{G}}\right]$, given a set of multi-view images $\{\mathrm{I}^v\}_{v=1}^V$, their corresponding image normal maps $\{\mathrm{N}^v\}_{v=1}^V$, and camera parameters $\{\mathrm{K}^v\}_{v=1}^V$, where V denotes the total number of views. The proposing model $g(\cdot)$ can be formally defined by the following equation:

$$s, q = g(\mathbf{x}; \Phi_{\mathbf{x}}(\{\mathrm{I}^v, \mathrm{N}^v, \mathrm{K}^v\}_{v=1}^V). \tag{1}$$

The occupancy value $s \in \{0, 1\}$ is a binary indicator signifying whether a point resides inside or outside the surface boundary of an individual, essentially distinguishing the geometric presence of the subject in the 3D space. The identification (ID) value $q \in \mathbb{R}$ provides a unique identifier to each point, allowing differentiation of individuals in close proximity by assigning distinct ID values.

3.2 Multi-View Local-Global Feature Module

The architecture of the $\Phi_{\mathbf{x}}$ module, which extracts local and global features $\left[\Phi_{\mathbf{x}}^{\mathrm{L}}, \Phi_{\mathbf{x}}^{\mathrm{G}}\right]$ from multi-view images, determines the quality of occupancy and ID fields. We introduce a local-global feature scheme through a dedicated feature extraction architecture.

Given a set of images I^v, normal maps N^v generated by the method described in IntegratedPIFu [5], and camera parameters K^v, our method begins by following PIFu [32] using an image encoder [25]. Each image and normal map extracted by [5] is processed individually within the multi-view inputs, allowing distinctive features to be captured from different perspectives. For any given query point $\mathbf{x}$ in 3D space, we project this point onto the 2D planes of the multi-view inputs to acquire pixel-aligned features. Formally, the feature extraction process can be described as follows:

$$F^v = f^{\mathrm{img}}(\mathrm{I}^v \oplus \mathrm{N}^v), \tag{2}$$

$$F_{\mathbf{x}}^v = \Pi(F^v, \mathrm{K}^v, \mathbf{x}), \quad \forall v \in \{1, 2, \ldots, V\}, \tag{3}$$

where F^v denotes the set of features extracted from concatenated v-th image and v-th normal map and $\oplus$ symbolizes channel-wise concatenate operation. f^{img} represents the image encoder function, designed to process each views.

$F_{\mathbf{x}}^v$ represents the features at the image pixel corresponding to the projection of point $\mathbf{x}$ onto the v-th image plane. The function $\Pi(\cdot)$ computes the location on the image plane where the 3D point $\mathbf{x}$ is projected and extracts the relevant features from F^v at that point. For the detailed implementation, refer to PIFu [32].

This approach ensures that the features $F_{\mathbf{x}}^v$ are aligned with the geometry of the scene as observed from multiple viewpoints, facilitating an accurate reconstruction of the 3D space.

The pixel-aligned features extracted from each view are then aggregated through a local-global process to create a comprehensive feature representation for each query point $\mathbf{x}$. This aggregation is performed using a view-to-view transformer encoder [38], formulated as:

$$\Phi^{\mathrm{L}}_{\mathbf{x}} = f^{\mathrm{e}}(\gamma(\mathbf{x}), F^1_{\mathbf{x}}, F^2_{\mathbf{x}}, \cdots, F^V_{\mathbf{x}}), \tag{4}$$

where $\Phi^{\mathrm{L}}_{\mathbf{x}}$ represents the local feature set obtained by fusing features across all views corresponding to the query point $\mathbf{x}$ with positional encoding $\gamma(\mathbf{x})$.

Inspired by the architectures of the Scene Representation Transformer (SRT) encoder [34] and decoder mechanism [36], the global feature module complements these local features with global scene context. The SRT encoder f^{se} receives the extracted features from all views along with their corresponding camera positions $\mathbf{o}$ and normalized ray direction $\mathbf{d}$ from each camera information to encapsulate global scene information into a compact representation $\mathbf{z}$. This scene representation serves as an input to the global feature decoder f^{sd}, which extracts global features Φ^{G}:

$$\mathbf{z} = f^{\mathrm{se}}(\{F^v, \mathbf{o}^v, \mathbf{d}^v\}_{v=1}^V), \tag{5}$$

$$\Phi^{\mathrm{G}}_x = f^{\mathrm{sd}}(\mathbf{z}, \Phi^{\mathrm{L}}_x), \tag{6}$$

where Φ^{G} represents the global features for multi-view images.

Finally, we forward the features $\Phi_{\mathbf{x}}(\cdot)$ into two multi-layer perceptrons (MLPs), ϕ^{s} and ϕ^{id} to retrieve occupancy values s and ID values q for each query point:

$$s = \phi^{\mathrm{s}}(\Phi^{\mathrm{L}}_x), \tag{7}$$

$$q = \phi^{\mathrm{id}}(\Phi^{\mathrm{G}}_x). \tag{8}$$

This method ensures a comprehensive feature set that enhances predictions by integrating both localized and globalized scene insights.

Detailed architectures of f^{e}, f^{se}, and f^{sd} are shown in section A.2 of the Appendix.

3.3 Training Objective

We train this model g with following objectives:

$$\mathcal{L} = \omega_s \mathcal{L}_{\text{MSE}} + \omega_{\text{contra}} \mathcal{L}_{\text{contra}} + \omega_{\text{group}} \mathcal{L}_{\text{group}}, \tag{9}$$

where $\mathcal{L}_{\text{MSE}}$ is used to train occupancy field, while $\mathcal{L}_{\text{contra}}$ and $\mathcal{L}_{\text{group}}$ focus on accurately identifying individual entities captured as ID fields.

For occupancy predictions, mean Squared Error (MSE) loss is defined for N query points as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (s_i - s_i^{\mathrm{gt}})^2, \tag{10}$$

where s_i is the predicted occupancy and s_i^{gt} is the ground truth for the i-th point. $\mathcal{L}_{\text{MSE}}$ reconstructs 3D geometries by minimizing the discrepancy between the predicted and actual occupancy values, crucial for capturing the intricate details of the scene.

To ensure that points associated with the same object are assigned identical predicted ID values, we tailor an associative embedding [24] consisting of contrastive loss [9, 12] and a grouping loss in the training objective. The contrastive loss is computed based on pairwise Euclidean distances among ID value, considering both positive pairs for the same instance label and negative pairs for a different instance label. For all query points $\{\mathbf{x}_i\}$, we have a set of predicted ID values $\{q_i\}$ and associated ground truth instance labels $\{l(\mathbf{x}_i)\}$ given from datasets. Then, the contrastive loss is formulated as:

$$\mathcal{L}_{\text{contra}} = \frac{\sum_{ij}(m_{ij}^{\text{pos}} \cdot d_{ij})}{P} + \frac{\sum_{ij}\left(m_{ij}^{\text{neg}} \cdot \max(0, -d_{ij} + \delta)\right)}{P}, \tag{11}$$

where $d_{ij} = ||q_i - q_j||$ is a pairwise Euclidean distances between ID values of i and j-th query points. $m_{ij}^{\text{pos}} = \mathbb{I}[l_i = l_j]$ and $m_{ij}^{\text{neg}} = \mathbb{I}[l_i \neq l_j]$ are mask value for indicating positive or negative pairs. δ is a predefined margin threshold, and P is the total number of pairs. $\mathbb{I}[\cdot]$ denotes an indicator function that returns 1 for true and returns 0 for false case. If $\delta = 1$, then the negative loss component is doubled. This function aims to minimize the distance between positive pairs while ensuring negative pairs are separated by at least the margin.

The grouping loss function is devised to refine the embeddings of predicted ID values by accomplishing two primary objectives: minimizing the variance within groups of values that correspond to the same ground truth label and maximizing the separation between groups linked to different labels. This function is articulated as follows:

$$\mathcal{L}_{\text{group}} = \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{l(\mathbf{x}) \in G_k} (x - \boldsymbol{\mu}_k)^2 + \sum_{k=1}^{K} \sum_{l=1, l \neq k}^{K} e^{-|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l|}. \tag{12}$$

In this equation, K denotes the total number of unique ID values present in the ground truth labels, signifying the distinct classifications for the predicted ID values. Each G_k represents the collection of predicted ID values that match the k-th unique ground truth label, forming groups of predictions intended to bear the same ID. The term $|G_k|$ reflects the size of each group, indicating the count of predictions it encompasses. The variable x refers to individually predicted ID value within the group G_k, and μ_i is the average of values in G_k, acting as the group's centroid. The notation $|\cdot|$ signifies the absolute value, ensuring that distances in the computations remain non-negative.

The first component of the grouping loss concentrates on reducing the squared distances between each predicted ID value x and its group's mean μ_k, thereby decreasing the variance within each group. Conversely, the second component introduces an exponential penalty on the closeness of centroids μ_k and μ_l from distinct groups, fostering clear separation between these groups by penalizing closely situated centroids. This bifocal approach of the loss function is crucial for steering the model towards generating cohesive yet distinctly separated embeddings, which is fundamental for the precise identification and differentiation of unique ID values.

3.4 Estimation of Interaction Geometry

To reconstruct the geometry of occupancy fields, we first retrieve the occupancy fields by querying dense points from the implicit fields g. After estimating the occupancy field, the geometry is reconstructed using the Marching Cubes algorithm [19]. Initially, we define the bounding box of the voxel grid. The Marching Cubes algorithm is then applied to the grid, where a surface is generated by interpolating the predicted occupancy values with $s > \tau$ within each voxel unit. Subsequently, the algorithm retrieves the ID values from the same voxel units and applies a color map, which facilitates the visual distinction of different individuals represented in the geometry. We use $\tau = 0.5$ for thresholding occupancy values.

The final contact fields can be estimated by the variance of predicted ID values. Close interaction among different instances typically leads to deficient information due to occluded images. Consequently, the uncertainty in the predicted geometry is increased when the information from multiple views is deficient. We choose the variance of predicted ID values among possible options as an uncertainty metric because Euclidean distance is adopted within the training objective for predicted ID values. For each voxel $\mathbf{x} = (x_1, x_2, x_3)$ in the 3D grid $\mathbf{V}$, we extract a local neighborhood $\mathcal{N}(\mathbf{x})$, excluding background value ($s < \tau$). We then calculate the standard deviation of the neighborhood values,

$$\sigma_{\mathbf{x}} = \sqrt{\frac{1}{|\mathcal{N}(\mathbf{x})|} \sum_{\mathbf{v} \in \mathcal{N}(\mathbf{x})} (\mathbf{v} - \boldsymbol{\mu})^2}, \tag{13}$$

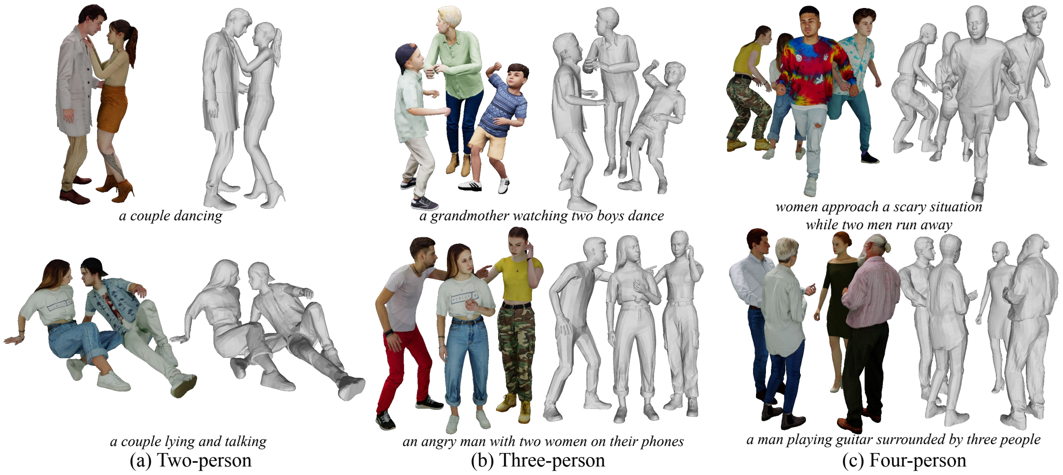
where μ is the mean of the neighborhood values and $|\mathcal{N}(\mathbf{x})|$ is the number of elements in the neighborhood.

A voxel is marked as a contact point if the standard deviation exceeds a threshold of $\tau_c = 0.25$. The formulation is as follows:

$$c(\mathbf{x}) = \begin{cases} \sigma_{\mathbf{x}} & \text{if } \sigma_{\mathbf{x}} > \tau_c \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

3.5 SynMPI: Synthetic Dataset for Multi-Person Interaction

Current multi-human benchmark datasets, such as Hi4D [42] and MultiHuman [43], are limited in size and scope, particularly in terms of the number of interacting individuals and the diversity of interaction scenarios. To address these limitations, we introduce a new synthetic dataset designed


a couple dancing
a grandmother watching two boys dance
women approach a scary situation while two men run away
a couple lying and talking
an angry man with two women on their phones
a man playing guitar surrounded by three people

(a) Two-person (b) Three-person (c) Four-person

Figure 2: Examples of multi-person interaction geometry in the SynMPI dataset. Each sample contains interactions involving (a) 2, (b) 3, or (c) 4 people. In each sample, the left side shows rendered RGB images and the right side shows rendered meshes. *Italic sentences* explain the multi-people interaction types of samples adopted from Character Creator 4 [31].

to encompass a broader range of interaction scenarios involving groups of up to four individuals. Figure 2 shows examples from our dataset.

Our dataset is multi-view and supports multi-person interactions, accommodating groups of two to four individuals to effectively capture the spatial dynamics of group interactions. We include elderly individuals and children, representing a wider range of identities for the enhanced diversity of dataset. Additionally, the synthetic data features individuals with varying ages, heights, weights, and clothing styles. We also incorporate multiple types of motions for each participant, enriching the dataset with a diverse set of dynamic interactions. The dataset encompasses approximately 50 distinct motion types, including dancing, running, and talking. The dataset includes 49 characters (26 female, 23 male) with 50 motion sequences across 43 scene configurations, totaling approximately 25,000 frames. For a detailed breakdown of the dataset, please refer to Appendix B.1.

4 Experiments

4.1 Datasets

We conduct our experiments on Hi4D [42] and the proposed SynMPI dataset, which include various multi-person interactive scenarios. Aggregated training splits of both Hi4D and SynMPI serve as the training set for our experiments. The test splits of each dataset serve as the test set.

Hi4D [42] This dataset offers samples which two individuals engage in various interactions, encompassing 20 subject pairs with diverse body shapes and appearances with 8 view images. It includes 3D human scans, instance segmentation masks at the vertices of the 3D scans and image pixels, SMPL [2] parameters, and contact information at the vertex level. We use a random split of 70% for training and 30% for evaluation.

SynMPI Our synthetic multi-person interaction dataset captures a wide range of interaction scenarios involving groups of more than two individuals, encompassing a diverse spectrum of dynamic interactions. From the our synthetic datasets, we use images from 8 views and 3D geometry for our experiments. We randomly split samples in SynMPI into 70% for training and 30% for evaluation.

4.2 Metrics

Following the evaluation protocols of existing studies [32, 43], we adopt four metrics to assess the quality of the interaction geometry. Chamfer Distance (CD) calculates the bidirectional disparity between points on the predicted and corresponding ground-truth mesh. Point to Surface (P2S) computes the unidirectional distance from each point of the ground-truth mesh to the nearest surface

Table 1: Evaluation results for multi-person interaction geometry. The values presented under the CP row indicate the threshold value, denoted by ϵ, which was employed to construct the pseudo G.T. contact map in 3D space for evaluation.

Model	Hi4D [42]							SynMPI (Ours)		
	CD↓	P2S↓	NC↑	CP↑				CD↓	P2S↓	NC↑
				0.025	0.05	0.075	0.1			
DMC [43]	0.631	0.495	0.768	-	-	-	-	0.804	0.800	0.688
Ours (w/o SRT)	0.468	0.402	0.888	0.317	0.424	0.458	0.492	0.630	0.492	0.827
Ours	**0.406**	**0.329**	**0.892**	**0.447**	**0.629**	**0.670**	**0.703**	**0.511**	**0.374**	**0.836**

Table 2: Ablation study on grouping loss function in Eq (12).

Ablation type	Squared distance	Exponential penalty	CD↓	P2S↓	NC↑	CP↑	
						0.05	0.075
(a)	not used	not used	0.462	0.363	0.892	0.111	0.187
(b)	used	not used	0.400	0.314	0.892	0.345	0.528
(c)	not used	used	0.532	0.403	0.880	0.228	0.335
(d)	used	used	0.406	0.329	0.892	0.629	0.670

on the corresponding predicted mesh based on the closest-set euclidean distances. Normal Consistency (NC) measures the difference of the normal vector between points on the predicted and corresponding ground-truth mesh with the nearest-set euclidean distances. Contact Precision (CP) is defined by the overlap between the estimated contact map and the pseudo ground truth contact map generated from ground truth meshes. For a detailed definition of the metrics, please refer to the Appendix A.4.

4.3 Baseline Models

DeepMultiCap (DMC) [43]: We compare our framework with existing methods [43] reconstructing 3D human with multi-view images. DMC leverages a 3D feature of SMPL mesh to infer information of occluded regions during the learning process of the pixel-aligned implicit function. For the integration of features from multiple views, it utilizes a transformer-based approach. They use 8-view images for their reconstruction process, and we followed the same setup. We use the public implementation of DMC [1] and apply LVD [10] on the SynMPI, as well as MVpose [11] on the Hi4D, to obtain SMPL parameters. We refer to Appendix B.2 for additional details.

4.4 Results and Analysis

Quantitative Results Reconstruction results are evaluated against the baseline method, as shown in Table 1. Our method demonstrated superior performance in terms of reconstruction quality metrics, indicating the effectiveness of our approach in accurately capturing and reconstructing 3D models of multiple people.

Qualitative Results Figure 3 illustrates the qualitative performance of our method in generating high-quality reconstructions of multiple people in close interaction scenarios. Compared to DMC [43], our method excels in handling dynamic interactions and heavy occlusion, which are common challenges in multi-person reconstruction tasks. DMC struggles with these scenarios, leading to less accurate SMPL estimations. For additional results, please refer to the appendix and supplementary video.

Ablation Study on Architecture We performed ablation studies to assess the impact of our proposed modules. Table 1 presents the performance of the geometry module without the global features. The results demonstrate that global features significantly enhance performance across all geometry-related metrics. Table 1 highlights the impact of global features on contact precision performance, demonstrating enhanced accuracy of contact predictions across a range of thresholds. Figure 5 visually compares contact precision performance with and without global features, illustrating substantial

[1] `https://github.com/DSaurus/DeepMultiCap`.

Table 3: Ablation on the number of views.

Model	# views	CD$\downarrow$	P2S$\downarrow$	NC$\uparrow$
DMC [43]	4	1.304	0.922	0.705
	8	0.631	0.495	0.768
Ours	4	0.761	0.472	0.870
	8	0.406	0.329	0.892

Table 4: Method for contact map estimation

Method	CP$\uparrow$	
	0.05	0.075
output meshes	0.518	0.621
variance estimation (ours)	0.629	0.670

improvements. Additionally, Figure 4 depicts the geometry performance with ID and ID field volume rendering, further demonstrating the positive impact of global features.

Ablation Study on Grouping Loss Grouping loss function defined in Eq (12) comprises two key terms: the first is the squared distance, and the second is the exponential penalty. In Table 2, model (a) is trained without the grouping loss, while model (d) is trained with the grouping loss. The exponential function in the second term encourages soft assignment to a specific instance or cluster. However, using only this term does not lead to improved grouping performance. We observe that the combination of both terms within the grouping loss function results in overall performance enhancement.

Contact Map Analysis Table 1 also presents the performance of our contact map using contact precision metrics. Our approach leverages unsupervised learning to predict contact fields in 3D space, eliminating the need for labeled training data, which is often difficult and costly to obtain. This is particularly advantageous in complex scenarios involving multiple individuals and interactions. The effectiveness of our method is significantly influenced by the resolution of the 3D data. High-resolution data provide detailed and dense information, enabling precise contact prediction. In contrast, low-resolution data can lead to less accurate results due to the sparse representation of the interactions, as illustrated in Figure 5. A key advantage of our approach is the use of an implicit contact field, which allows for flexible changes in resolution. This flexibility enables our method to adapt to various data resolutions without compromising the integrity of the contact prediction. Thus, our method excels in flexibility and reduces dependency on extensively labeled datasets. However, ensuring adequate resolution of 3D data is crucial for achieving optimal accuracy and reliability in contact field estimation. In this paper, we use a resolution of 256^3 to estimate the contact field.

To further assess our contact predictions, we directly infer contacts from the output instance meshes by examining geometric proximity and the surface identifiers used to generate the contact map from the pseudo ground truth instance meshes.Table 4 presents the results for both contact map estimation methods. Our variance-based estimation of the contact field in 3D space yields better results than the mesh-based inference. This is because the variance estimation method benefits from high-resolution 3D data and operates in continuous space, allowing for precise localization of contact areas without intermediary steps that could introduce errors. In contrast, the mesh-based inference relies on reconstructed meshes that may lack fine details due to resolution limitations or reconstruction errors, leading to less accurate contact predictions. Additional details are provided in Appendix B.3.

Ablation Study on the Number of Views Table 3 presents an ablation study on the number of views. Our method consistently outperforms DMC across all metrics—Chamfer Distance (CD), Point-to-Surface Distance (P2S), and Normal Consistency (NC)—in both the 4-view and 8-view settings.

5 Conclusion

This paper addresses the intricate challenges associated with the 3D reconstruction of multiple interacting human bodies in close proximity, an area critical for applications in virtual reality, augmented reality, robotics, and surveillance. Our approach overcomes the limitations of traditional methods that rely on models like the Skinned Multi-Person Linear (SMPL), which often struggle in scenarios with dynamic interactions and occlusions. By employing advanced methodologies, including a multi-view feature transformer and a global scene feature extraction transformer, our approach not only preserves the unique identities and spatial information of each individual but also enhances the accuracy of 3D reconstructions.

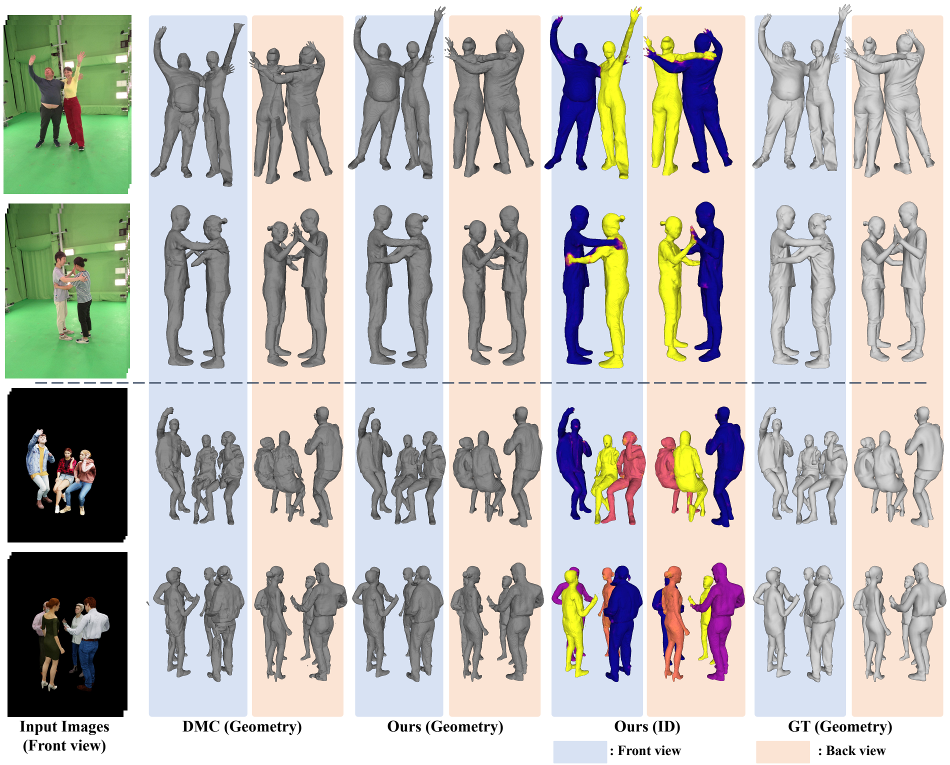


Figure 3: We compare our method to baseline DMC [43] on Hi4D (top) and SynMPI (bottom) test set. From left to right columns, we show the input multi-view images, the generated geometry by each method, and ground truth (GT) Geometry.

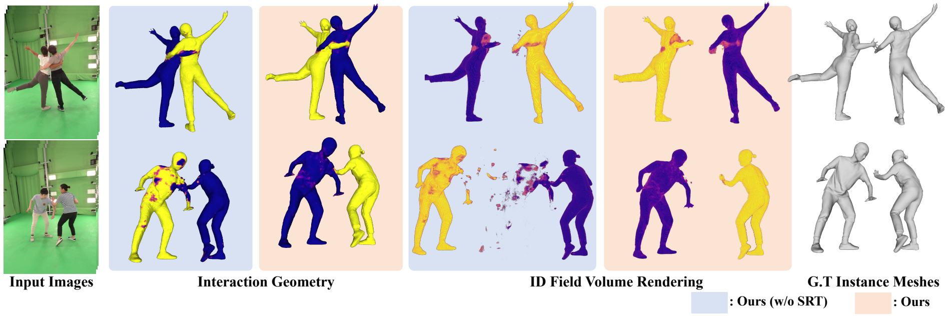


Figure 4: Visualization of reconstructed multi-person interaction geometry and instance-wise volume rendering with ID fields for visualizing occluded regions during interaction.

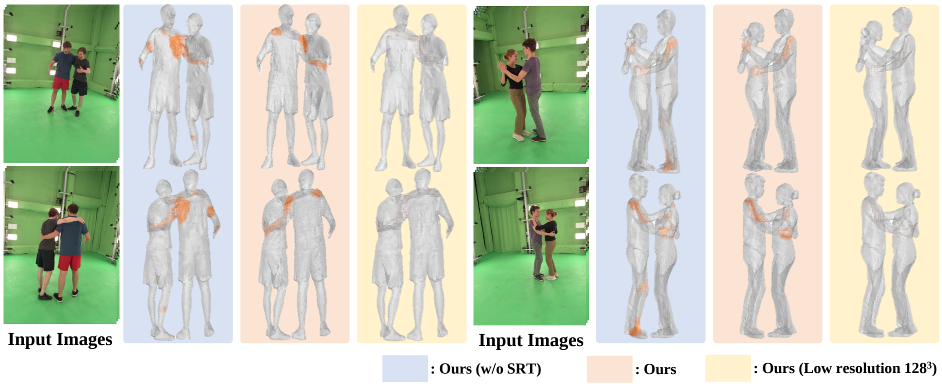


Figure 5: Comparison of the effect of global features and 3D resolution on estimated contact fields. Ours (w/o global) excludes global features, Ours (low resolution 128^3) uses low resolution.

Acknowledgments and Disclosure of Funding

This work was supported by Korea Creative Content Agency (KOCCA) grant funded by the Korea government (MCST) (No. RS-2023-00224427, Development of intelligent virtual sports service technology for breaking dance).

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. "Photorealistic monocular 3d reconstruction of humans wearing clothing". In: *CVPR*. 2022.

[2] 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*. 2016.

[3] Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jaeshin Yoon, and Seungryul Baek. "3D Reconstruction of Interacting Multi-Person in Clothing From a Single Image". In: *WACV*. 2024.

[4] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. "Multi-Person 3D Pose and Shape Estimation via Inverse Kinematics and Refinement". In: *ECCV*. 2022.

[5] Kennard Yanting Chan, Guosheng Lin, Haiyu Zhao, and Weisi Lin. "Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction". In: *ECCV*. 2022.

[6] Dongyue Chen, Yuanyuan Song, Fangzheng Liang, Teng Ma, Xiaoming Zhu, and Tong Jia. "3D human body reconstruction based on SMPL model". In: *The Visual Computer* (2023).

[7] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. "gDNA: Towards Generative Detailed Neural Avatars". In: *CVPR*. 2022.

[8] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. "Learning to estimate robust 3D human mesh from in-the-wild crowded scenes". In: *CVPR*. 2022.

[9] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *CVPR*. 2005.

[10] Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. "Learned vertex descent: A new direction for 3d human model fitting". In: *ECCV*. 2022.

[11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. "Fast and robust multi-person 3d pose estimation from multiple views". In: *CVPR*. 2019.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality Reduction by Learning an Invariant Mapping". In: *CVPR*. 2006.

[13] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. "Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction". In: *NeurIPS*. 2020.

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. "End-to-End Recovery of Human Shape and Pose". In: *CVPR*. 2018.

[15] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. "Convolutional Mesh Regression for Single-Image Human Shape Reconstruction". In: *CVPR*. 2019.

[16] Jungeun Lee, Sanghun Kim, Hansol Lee, Tserendorj Adiya, and Hwasup Lim. "PIDiffu: Pixel-Aligned Diffusion Model for High-Fidelity Clothed Human Reconstruction". In: *WACV*. 2024.

[17] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* (1982).

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. "SMPL: A skinned multi-person linear model". In: *TOG* (2015).

[19] William E. Lorensen and Author PictureHarvey E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm". In: *Computer graphics* (1987).

[20] Yang Lu, Han Yu, Wei Ni, and Liang Song. "3D real-time human reconstruction with a single RGBD camera". In: *Applied Intelligence* (2023).

[21] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. "Occupancy networks: Learning 3d reconstruction in function space". In: *CVPR*. 2019.

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* (2021).

[23] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. “Multi-person implicit reconstruction from a single image”. In: *CVPR*. 2021.

[24] Alejandro Newell, Zhiao Huang, and Jia Deng. “Associative Embedding: End-to-End Learning for Joint Detection and Grouping”. In: *CVPR*. 2016.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *ECCV*. 2016.

[26] NVIDIA. *Kaolin*. `https://github.com/NVIDIAGameWorks/kaolin/`.

[27] NVIDIA. *Omniverse USD Composer*. `https://docs.omniverse.nvidia.com/composer/latest/index.html/`.

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. “Deepsdf: Learning continuous signed distance functions for shape representation”. In: *CVPR*. 2019.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. “Automatic differentiation in PyTorch”. In: *NeurIPS 2017 Workshop on Autodiff*. 2017.

[30] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. “Learning to estimate 3D human pose and shape from a single color image”. In: *CVPR*. 2018.

[31] Reallusion. *CharacterCreator4*. `https://www.reallusion.com/character-creator/`.

[32] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization”. In: *ICCV*. 2019.

[33] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. “Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization”. In: *CVPR*. 2020.

[34] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. “Object scene representation transformer”. In: *NeurIPS*. 2022.

[35] Mehdi SM Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. “Rust: Latent neural scene representations from unposed imagery”. In: *CVPR*. 2023.

[36] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. “Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations”. In: *CVPR*. 2022.

[37] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. “Probabilistic 3D Human Shape and Pose Estimation From Multiple Unconstrained Images in the Wild”. In: *CVPR*. 2021.

[38] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. “Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering”. In: *CVPR*. 2022.

[39] 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*. 2022.

[40] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. “Direct Multi-view Multi-person 3D Pose Estimation”. In: *NeurIPS*. 2021.

[41] Chenyan Wu, Yandong Li, Xianfeng Tang, and James Wang. “MUG: Multi-human graph network for 3D mesh reconstruction from 2D pose”. In: *arXiv preprint arXiv:2205.12583* (2022).

[42] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. “Hi4d: 4d instance segmentation of close human interaction”. In: *CVPR*. 2023.

[43] 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*. 2021.

[44] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. “DeepHuman: 3D Human Reconstruction From a Single Image”. In: *ICCV*. 2019.

A Implementation Details

A.1 Training

We trained our model on the Hi4D [42] dataset and our dataset SynMPI. Following the PIFu [32] process to extract pixel-aligned features, we sampled 6000 points used during training. The training was conducted with a batch size of 4 and a learning rate of $1e-4$. Our learning rate schedule involved decaying the initial learning rate by a specified factor (gamma) at predetermined epochs, as defined in our schedule. The model was optimized using the RMSprop optimizer. In our experiments, we set $\omega_s = 1$, $\omega_{\text{contra}} = 0.1$, and $\omega_{\text{group}} = 0.1$ for the weights of loss functions. Implemented using PyTorch [29], the entire training process spanned approximately two days and covered 100 epochs using two NVIDIA A100 GPUs. Inference of each instance requires around 60 seconds on the same GPU.

A.2 Architecture

Multi-View Feature Fusion Module $f^{\mathbf{e}}$. Our objective with f^{e} is to effectively aggregate features from multiple view inputs. Utilizing a view-to-view transformer architecture inspired by Double-Field [38], we process pixel-aligned features to facilitate this aggregation. The operation of our transformer is mathematically represented as:

$$\begin{aligned} Q_{\mathbf{x}}^{f}, K_{\mathbf{x}}^{f}, V_{\mathbf{x}}^{f} &= \varepsilon_{K,Q,V}^{f}(F_{\mathbf{x}}^{1}, F_{\mathbf{x}}^{2}, ..., F_{\mathbf{x}}^{V,}), \\ \Phi_{\mathbf{x}}^{L} &= \varepsilon^{f}(\text{Attention}(Q_{\mathbf{x}}^{f}, K_{\mathbf{x}}^{f}, V_{\mathbf{x}}^{f})), \end{aligned} \tag{15}$$

where $Q_{\mathbf{x}}^{f}$, $K_{\mathbf{x}}^{f}$, and $V_{\mathbf{x}}^{f}$ denote the query, key, and value matrices generated from the input features, respectively. Following self-attention, the features are further refined through feed-forward networks f^{F} to obtain the local feature set $\Phi_{\mathbf{x}}^{L}$.

SRT Encoder $f^{\mathbf{se}}$. SRT Encoder, denoted as f^{se}, aims to encapsulate 3D scene information into a comprehensive set-latent scene representation z. Following methodologies similar to those described by Object Scene Representation Transformer (OSRT) [34], our encoder leverages the Transformer's self-attention mechanism to aggregate spatial and feature information from multiple views into a single scene representation. This process is formalized as:

$$\{F_v\}_{v=1}^{V} = \varepsilon_f(\{F^v, \varepsilon_{\text{ray}}(\mathbf{o}^v, \mathbf{d}^v)\}_{v=1}^{V}), \tag{16}$$

where ε_f and ε_{ray} represents the conv block yielding a set of features $\{F_v\}_{v=1}^{V}$. Subsequently, these features are aggregated into a set of flatted patch embeddings $\{E_i\}_{i=1}^{N}$, where N denotes the total number of patches across all images. This aggregation can be mathematically represented as:

$$\{E_i\}_{i=1}^{N} = \varepsilon_{\text{patch}}(\{F_v\}_{v=1}^{V}), \tag{17}$$

The transformer encoder, $\mathcal{T}^{e}$, then processes this set of embeddings to generate the final set-latent scene representation z:

$$z = \mathcal{T}^{e}(\{E_i\}_{i=1}^{N}), \tag{18}$$

where z fully encapsulates the observed 3D scene, encoding comprehensive spatial and visual scene information. It is imperative to note that set-latent scene representation z embodies the comprehensive understanding of the specific 3D scene as observed through the corresponding set of images. This representation, characterized by its ability to maintain the integrity and richness of scene's spatial and feature information, is pivotal for subsequent reconstruction and analysis tasks.

Global Feature Decoder $f^{\mathbf{sd}}$. We employ SRT decoder to extract global features from scene representation z. There are some differences with original SRT decoder [36]. The main difference is that local feature $\Phi_{\mathbf{x}}^{\mathrm{L}}$ is used for query and value in multi-head attention mechanism. This modification ensures that:

$$\Phi_{x}^{\mathrm{G}} = f^{sd}(z, \Phi_{\mathbf{x}}^{\mathrm{L}}), \tag{19}$$

where Φ_{x}^{G} represents the globally decoded feature. This enables the decoder to dynamically focus on relevant scene information, thereby facilitating detailed and accurate 3D reconstructions and precise occupancy and identification predictions.

A.3 Dataset Construction

To create our dataset, we first acquired characters of various ages and interaction motion sequences from Character Creator 4 [31]. We then composed scenes featuring multiple characters using Omniverse USD Composer [27]. To facilitate the dataset generation process, we modified the Kaolin rendering tool [26] to include tasks such as multi-person normalization, enabling us to achieve the desired outputs. This process allows us to generate multi-view rendered images, mask images, instance masks, and 3D geometry.

A.4 Evaluation Metrics

The definition of evaluation metrics are shown in below. $\mathcal{P}$ and $\mathcal{Q}$ refer to the set of 3D points.

Chamfer Distance (CD) This metric calculates the bidirectional disparity between points on the predicted and corresponding ground-truth mesh. It computes the euclidean distance from each point to nearest surface on other mesh. Lower value of CD metric indicates a higher fidelity of reconstruction.

$$\mathrm{CD}(\mathcal{P}, \mathcal{Q}) = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \min_{\mathbf{q} \in \mathcal{Q}} \|\mathbf{p} - \mathbf{q}\|^2 + \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{q} \in \mathcal{Q}} \min_{\mathbf{p} \in \mathcal{P}} \|\mathbf{q} - \mathbf{p}\|^2 \tag{20}$$

Point to Surface (P2S) This metric computes the unidirectional distance from each point of ground-truth mesh to the nearest surface on the corresponding predicted mesh based on the closest-set euclidean distances. Lower value of P2S metric means superior reconstruction accuracy.

$$\mathrm{P2S}(\mathcal{P}, \mathcal{Q}) = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \min_{\mathbf{q} \in \mathcal{Q}} \|\mathbf{p} - \mathbf{q}\|^2 \tag{21}$$

Normal Consistency (NC) This metric computes the difference of normal vector between points on the predicted and corresponding ground-truth mesh with the nearest-set euclidean distances. It uses the bidirectional way for calculating the difference. Lower value of NC metric indicates a higher fidelity of reconstruction.

$$\mathrm{NC}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \left(1 - \mathbf{n}_{\mathbf{p}} \cdot \mathbf{n}_{\text{nearest}(\mathbf{p}, \mathcal{Q})}\right) + \frac{1}{2|\mathcal{Q}|} \sum_{\mathbf{q} \in \mathcal{Q}} \left(1 - \mathbf{n}_{\mathbf{q}} \cdot \mathbf{n}_{\text{nearest}(\mathbf{q}, \mathcal{P})}\right) \tag{22}$$

Contact Precision (CP) We first identify sample points in the voxel grid using the ground truth mesh $\mathcal{M}_{\text{gt}}$,

$$\mathbf{P}_{\text{inside}} = \{\mathbf{p} \in \mathbf{P} \mid \mathcal{M}_{\text{gt}} \text{ contains } \mathbf{p}\}. \tag{23}$$

For each point $\mathbf{p}_i \in \mathbf{P}_{\text{inside}}$, we determine the nearest points on the mesh surface $\mathcal{M}$ and assign surface identifiers s_i based on face indices f_i. Using a KD-Tree, we find neighboring points within a specified distance threshold ϵ,

$$N_i = \{\mathbf{p}_j \mid \|\mathbf{p}_i - \mathbf{p}_j\| < \epsilon \text{ and } j \neq i\}. \tag{24}$$

A contact is marked if any neighboring point has a different surface identifier,

$$\text{contact}_i = \begin{cases} 1 & \text{if } \exists \mathbf{p}_j \in N_i \text{ such that } s_j \neq s_i \\ 0 & \text{otherwise} \end{cases}. \tag{25}$$

This process results in contact labels that are contextually relevant to the mesh's surface features, enabling validation of our contact prediction algorithms using precision as the evaluation metric.

The contact precision is defined by the overlap between the estimated contact map $\mathbf{E}$ and the pseudo ground truth contact map $\mathbf{T}$ generated from ground truth meshes. Then, the precision is given by

$$P(\mathbf{T}, \mathbf{E}) = \frac{|\mathbf{T} \cap \mathbf{E}|}{|\mathbf{E}|}, \tag{26}$$

where $|\mathbf{T} \cap \mathbf{E}|$ is the number of true positives (correctly predicted contact points), and $|\mathbf{E}|$ is the total number of predicted contact points. A higher precision value indicates that a greater proportion of the contact points predicted by the model are correct, signifying fewer false positives. This metric is crucial for assessing the accuracy of our contact prediction algorithms, ensuring that the predicted contacts closely align with the contacts defined by the pseudo ground truth.

Table A: Evaluation excluding SynMPI dataset on Hi4D test sets

	Geometry			Contact Precision ↑			
	CD↓	P2S↓	NC↑	0.025	0.05	0.75	0.1
w/o synthetic data	0.499	0.418	0.885	0.351	0.482	0.514	0.542
w/ synthetic data	**0.406**	**0.329**	**0.892**	**0.447**	**0.629**	**0.670**	**0.703**

Table B: Ablation study on SPML initialization method for synthetic datasets in DMC

Model	SMPL method	CD↓	P2S↓	NC↑
DMC	MVPose [11]	0.805	0.489	0.771
DMC	LVD [10]	0.631	0.495	0.768

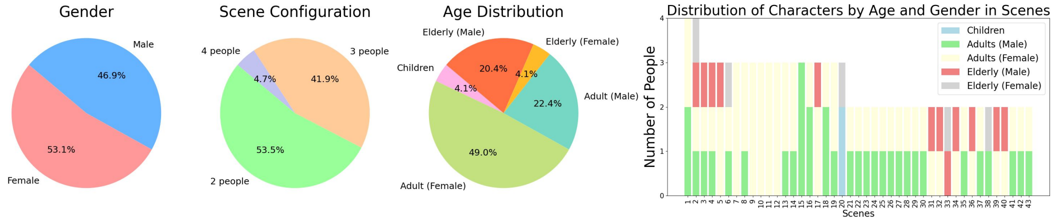


Figure A: Statistics of the SynMPI dataset.

B Analysis

B.1 Datasets

Table A presents the results of an ablation study using our SynMPI dataset. Training models with SynMPI alongside Hi4D [42] led to improved performance on the Hi4D test set. This improvement is mainly due to the large diversity within our dataset, which includes variations in age, gender, and scene composition, as shown in Figure A.

B.2 Baseline Models

For the baseline models discussed in Section 4.3, we employed different SMPL [18] acquisition methods tailored to each dataset. Specifically, for the HI4D dataset, we used MVPose [11], following the experimental setup described in the HI4D paper, where DMC [43] was run using MVPose for SMPL acquisition. We adopted this approach to ensure consistency and comparability.

Table B presents the ablation study on initial SMPL methods. However, we encountered challenges when using MVPose for our synthetic dataset, as its modules were trained on real data, leading to less accurate SMPL estimations on synthetic data. To address this, we utilized Learned Vertex Descent (LVD)[10], which is designed to fit SMPL to a 3D human model (3D scan) and has proven to provide more accurate results in this context. Although LVD is originally intended for single-person scenarios and may be sensitive to occlusions, we selected it for the synthetic dataset to achieve accurate results in our study.

B.3 Contact map

We also measure contact precision using the generated instance meshes as well as our proposed variance estimation of contact fields. To provide further insights into our method, we explain how different individuals are distinguished using predicted ID values during mesh generation.

To generate instance meshes from our implicit fields, our algorithm identifies and marks regions of interest based on the occupancy field during inference, excluding the background. This process generates both an ID field and a contact field. The normalized ID values within these regions

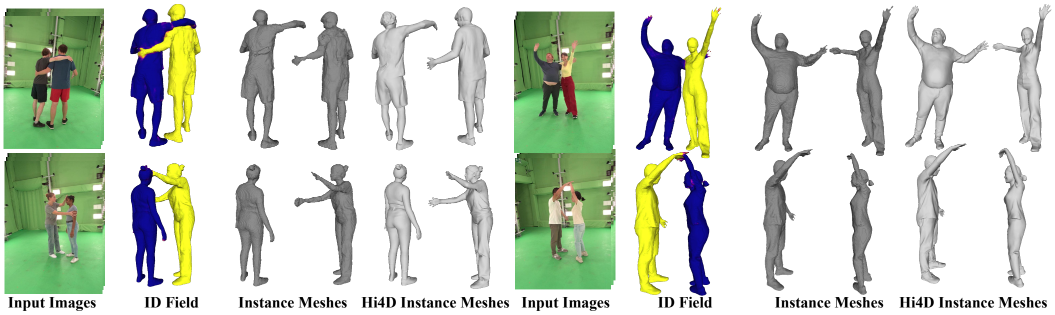


Figure B: Example of Instance meshes.

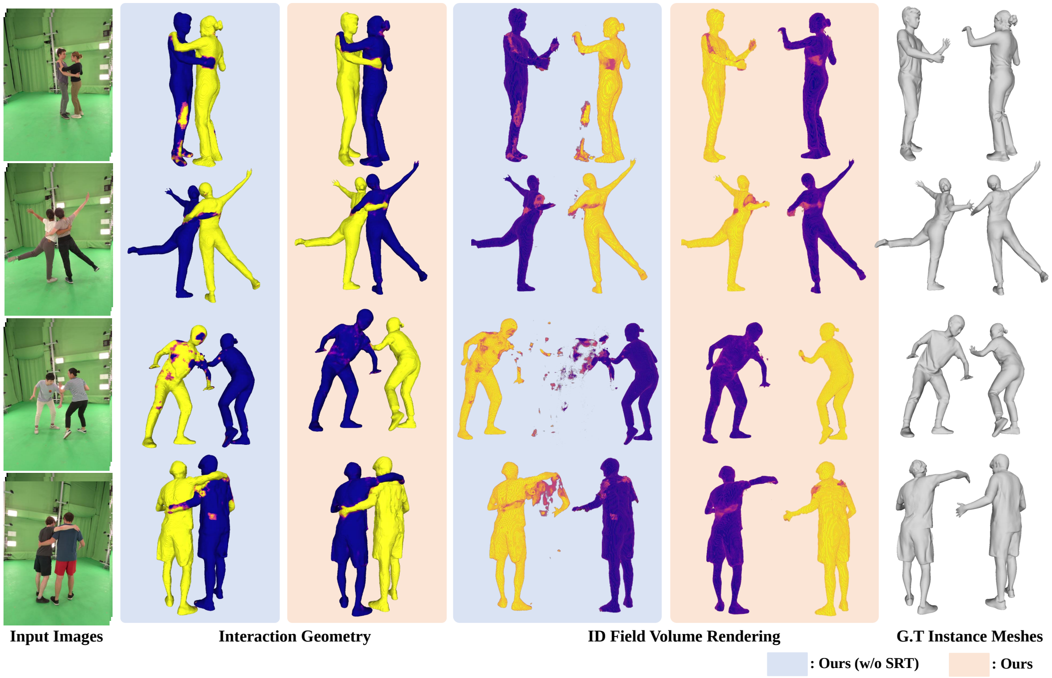


Figure C: Visualization of reconstructed multi-person interaction geometry and instance-wise volume rendering with ID fields for visualizing occluded regions during interaction.

are processed using k-Means clustering [17], grouping the data into clusters, with each cluster representing a different individual and including the associated contact regions.

After clustering, each cluster is isolated with a binary mask, which is smoothed using a Gaussian filter to create a blending mask that ensures smooth transitions at boundaries. This blending mask is applied to the occupancy field to enhance boundary details. Finally, the marching cubes algorithm generates a 3D mesh for each cluster from the processed occupancy field. These steps allow us to reconstruct instance meshes that support further evaluation of contact predictions.

Figure B illustrates the results of our instance meshes.

C Results

C.1 Visualization

We present additional visualization result samples in Figures D, E, G, and H. Additionally, Figure C shows further visualization results from our ablation study on architecture. Figure F provides more examples comparing contact maps based on 3D space resolution.

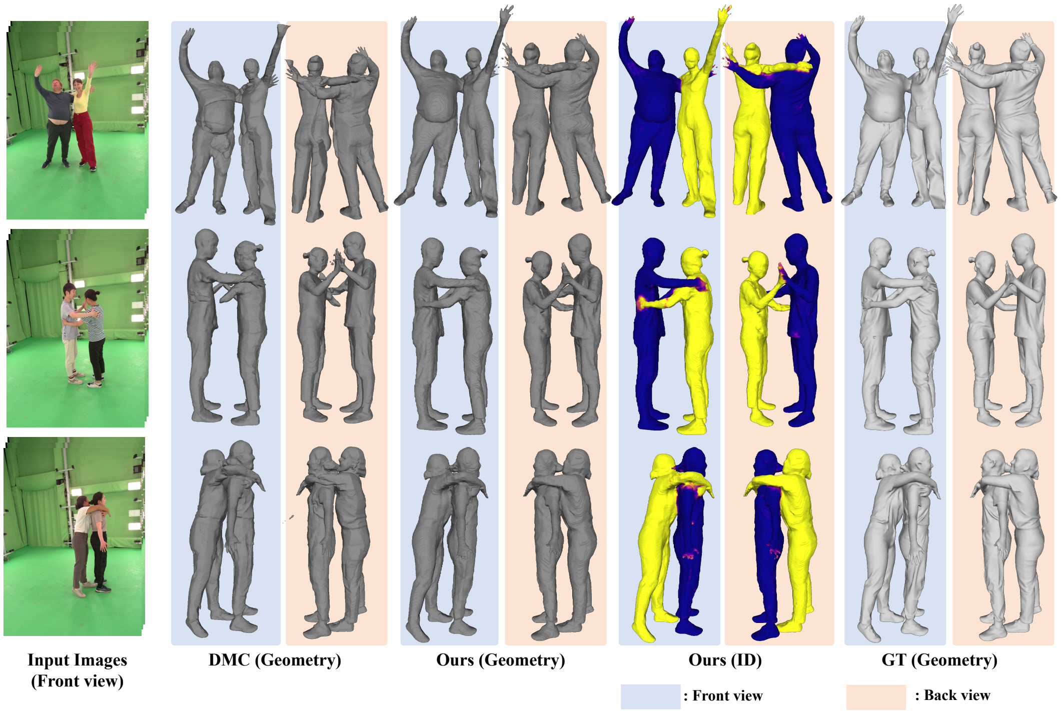


Figure D: Extended visualization of our method compared to baseline DMC [43] on Hi4D test split.

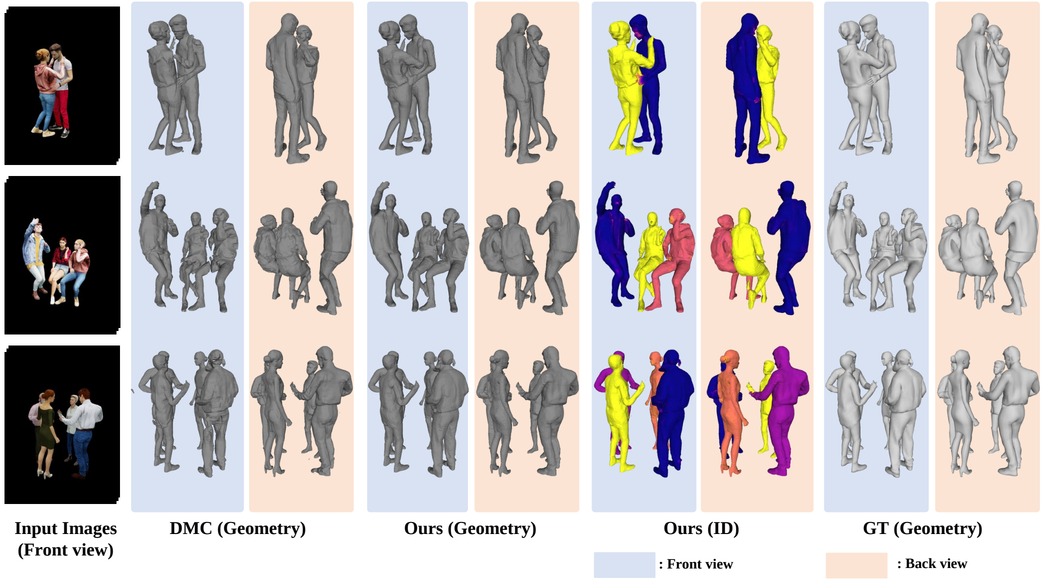


Figure E: Extended visualization of our method compared to baseline DMC [43] on test split of our synthetic dataset.

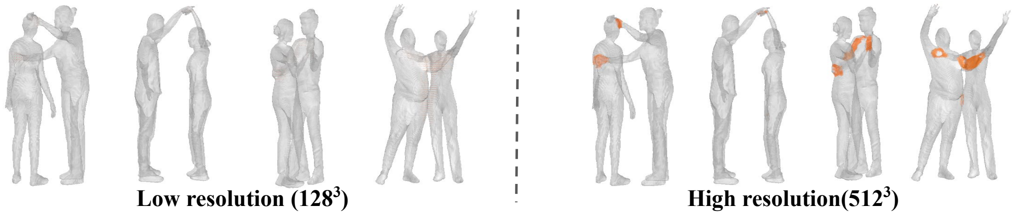


Figure F: Example of low resolution contact maps.

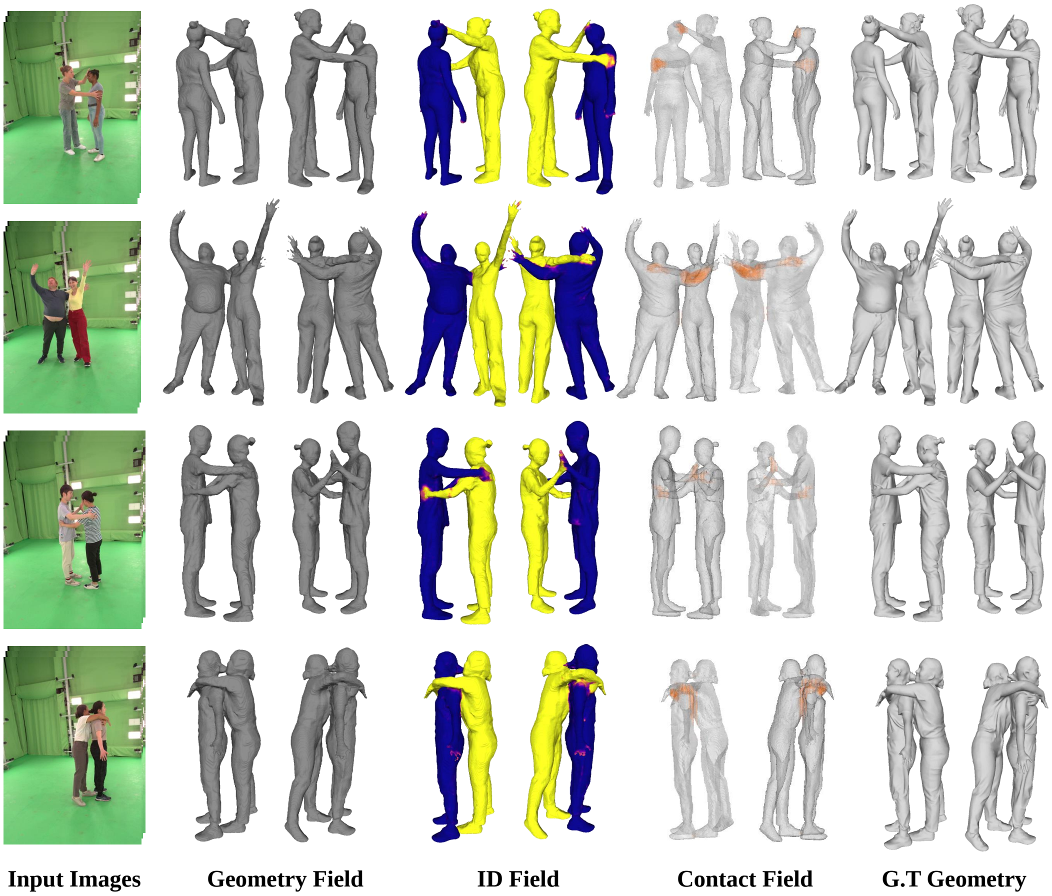


Figure G: Contact Field results of ours in Figure. 3

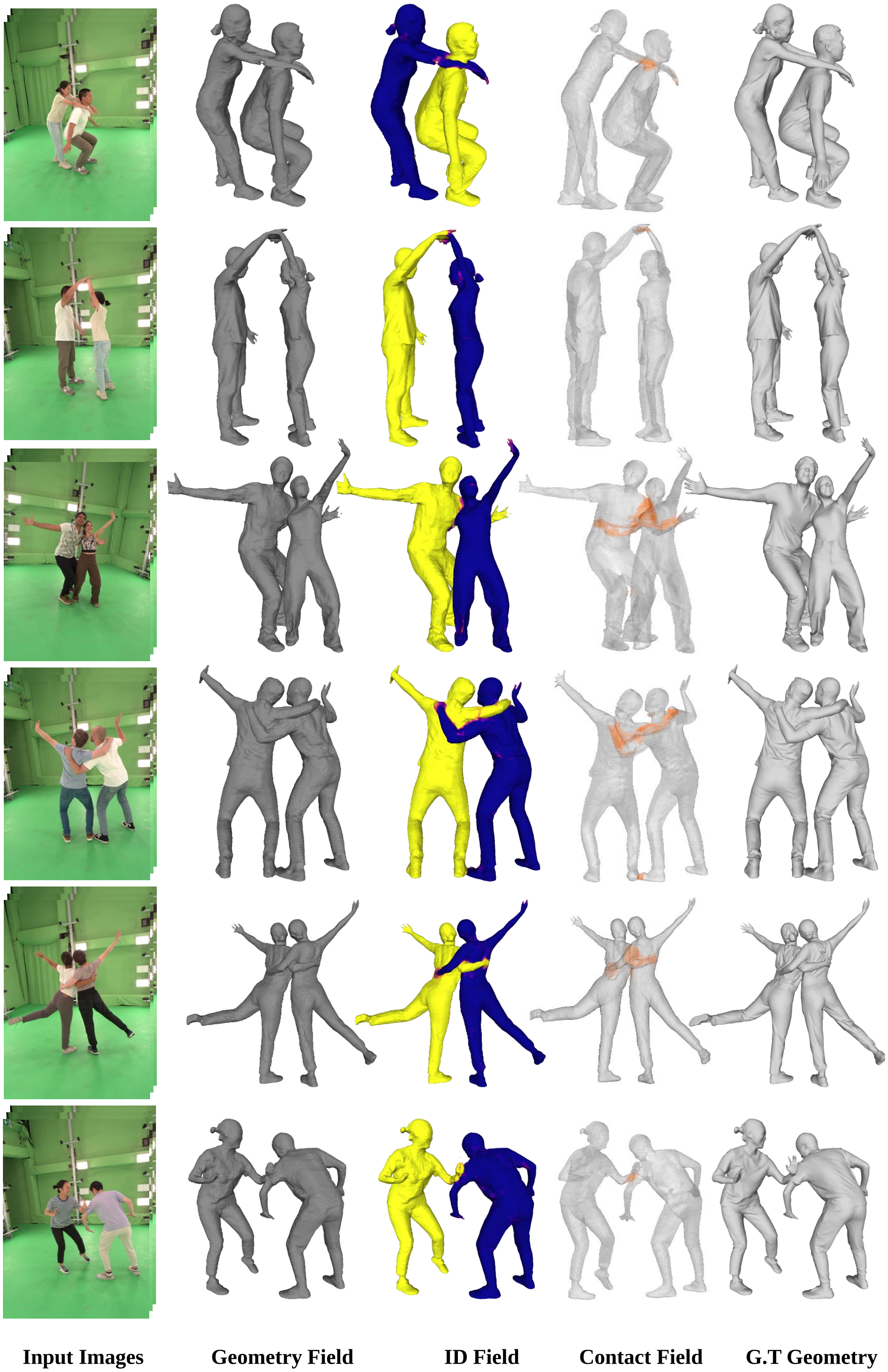


Figure H: Additional results of our method ContactField in Hi4D test split.

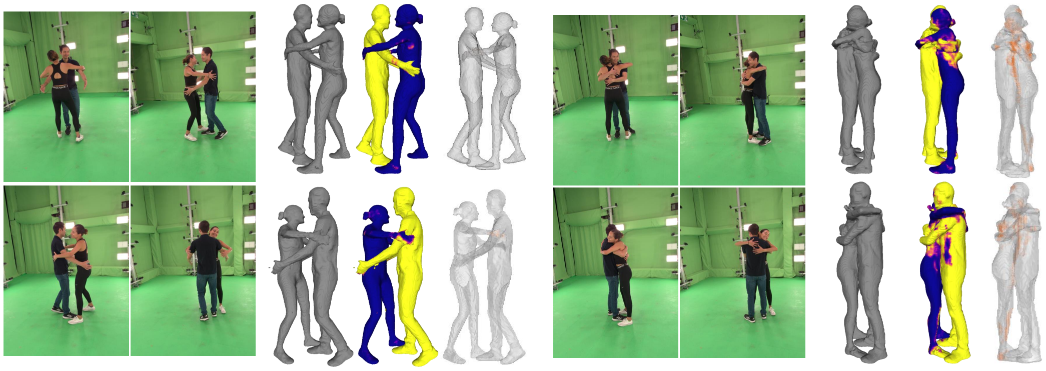


Figure I: Failure case.

D Discussion

D.1 Limitation

Despite its effectiveness, our approach introduces certain limitations, such as resolution constraints when reconstructing all elements simultaneously, which can affect the finer details of the models. Especially, we present failure cases in (d) of Figure I. Since we do not incorporate spatial prior such as SMPL, predicting identity (ID) in extreme poses, such as hugging, becomes challenging. However, as shown in (b) from a few frames earlier, the prediction is accurate. This suggests that incorporating a temporal module could be a promising direction for future work. Still, experimental results affirm the superiority of our approach over traditional methods, indicating significant potential for future enhancements in complex interaction scenarios and larger group dynamics. Moving forward, we aim to refine our techniques to address these resolution limitations and explore broader applications, further advancing the realism and functionality of 3D human body reconstructions.

E Broader Impacts

This novel implicit field representation for multi-person interaction geometry in 3D spaces has the potential to advance various applications in healthcare, sports, and security by enabling more accurate and detailed reconstructions of human interactions. However, the enhanced ability to capture such interactions also raises important concerns regarding privacy and ethical use. To mitigate these risks, it is essential to implement robust data protection measures, establish clear ethical guidelines, and ensure compliance with privacy laws.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. We explain our motivation and contribution of our work in the main paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explain the limitation of the proposed method in section D.1 of appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work is empirical work on interaction geometry for multi-person interaction, which do not requires toehry and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explain training details in section A of the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not release the code or data because our work deals with human subject. Although we can not release the code, we will provide the detailed description for our implementation to reproduce all the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain both training and architecture details in section A of the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We do not present any statistical significance in the table or figure due to strong supervision in reconstruction task. We describe all the details about the experiment for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computing resources we used for experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We explain broader impacts of the current work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release the code or data because our work deals with human subject.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We describe all the datasets and models we used in the manuscript.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We describe the details about new synthetic dataset we introduced in the manuscript.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We use the publicly available dataset and synthetic dataset for human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We use the publicly available dataset and synthetic dataset for human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.