DOI: 10.1111/cgf.14982
Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

Reconstructing 3D Human Pose from RGB-D Data with Occlusions

Bowen Dang[1] Xi Zhao[†1] Bowen Zhang[1] He Wang[2]

[1]Xi'an Jiaotong University, China [2]University College London, United Kingdom

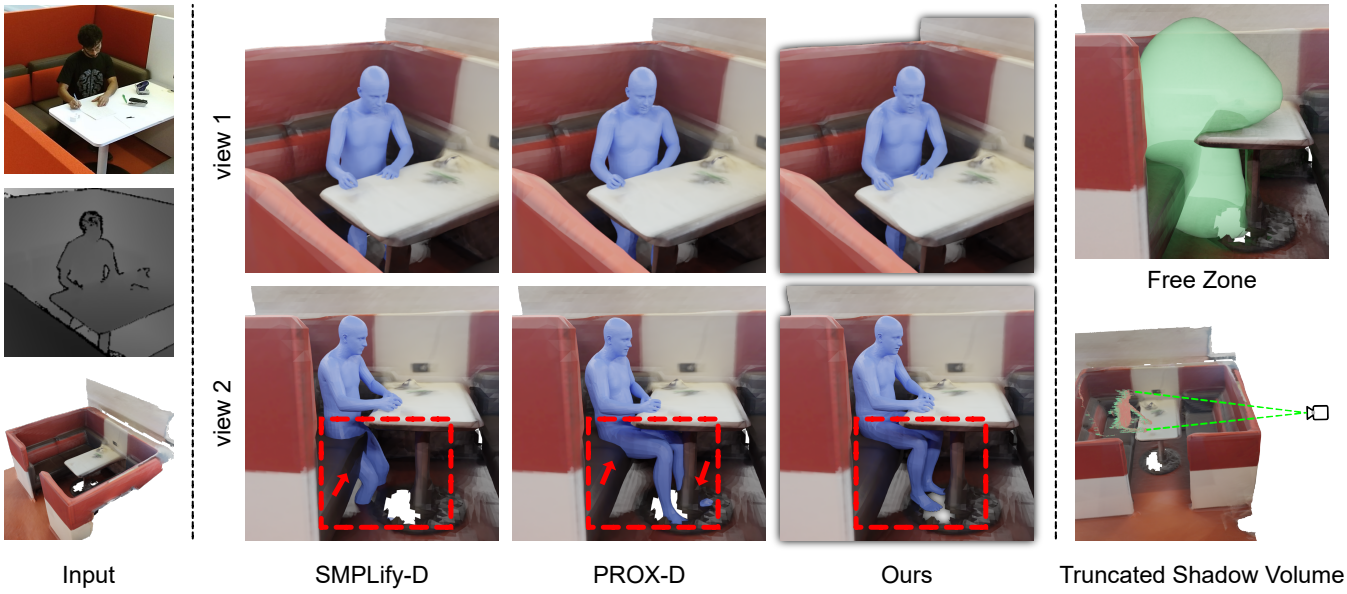

Figure 1: *Given a monocular RGB-D image and the scene mesh, our method reconstructs the 3D human body with a more plausible pose compared with SMPLify-D and PROX-D [HCTB19] by reducing the solution space using free zone and truncated shadow volume. The difference between results is highlighted using red dashed box and arrows.*

Abstract

We propose a new method to reconstruct the 3D human body from RGB-D images with occlusions. The foremost challenge is the incompleteness of the RGB-D data due to occlusions between the body and the environment, leading to implausible reconstructions that suffer from severe human-scene penetration. To reconstruct a semantically and physically plausible human body, we propose to reduce the solution space based on scene information and prior knowledge. Our key idea is to constrain the solution space of the human body by considering the occluded body parts and visible body parts separately: modeling all plausible poses where the occluded body parts do not penetrate the scene, and constraining the visible body parts using depth data. Specifically, the first component is realized by a neural network that estimates the candidate region named the "free zone", a region carved out of the open space within which it is safe to search for poses of the invisible body parts without concern for penetration. The second component constrains the visible body parts using the "truncated shadow volume" of the scanned body point cloud. Furthermore, we propose to use a volume matching strategy, which yields better performance than surface matching, to match the human body with the confined region. We conducted experiments on the PROX dataset, and the results demonstrate that our method produces more accurate and plausible results compared with other methods.

CCS Concepts

• ***Computing methodologies*** → ***Shape modeling; Reconstruction;***

† Corresponding Author

1. Introduction

3D human reconstruction is an important research area with broad applications in human behavior understanding and human-scene interaction analysis [ZZB*21]. Most current works focus on reconstructing 3D human body from monocular RGB images and matching the reconstructed result with the 2D image [BKL*16, KBJM18, PCG*19, KPBD19, JNV21, LWL21, CPMAMN22]. With the development of human-scene interaction datasets [SCH*16, HCTB19, WCR*19, ZMZ*22], many methods [HCTB19, ZZB*21, LSS*22] have been dedicated to reconstructing 3D human body that not only matches with the 2D human contour in image but also keeps reasonable spatial relationship with the environment by using 3D scene information. In this paper, we also follow this line and work on 3D human reconstruction from monocular RGB-D images. We specifically focus on situations which contain close interactions and serious occlusions between the human and the environment.

3D human reconstruction methods can be divided into two types: optimization-based and regression-based methods. Optimization-based methods attempt to reconstruct the 3D human body by minimizing an objective function to optimize the parameters of the human body model [BKL*16, PCG*19] or vertices [CPMAMN22]. Regression-based methods directly regress the parameters of the human body model [KBJM18] or vertices [LWL21] in an end-to-end manner, which requires a lot of data to train and might lead to inaccurate pose. So the regression-based method is normally followed by a post-process to optimize the pose [KPBD19, JNV21]. Although Current methods all consider the penetration terms [HCTB19], they suffer from penetrations between the body and the environment. The most important reason is that it is quite hard to improve or resolve the penetration once it happens and the system may stuck in the local minimum. The penetration problem is getting even harder when dealing with scenes with close interactions and serious occlusions.

Our method falls into the optimization type. Our key idea is to explicitly reduce the solution space based on scene information and prior knowledge. We propose two strategies to constrain the solution space. First, with the 3D scene around the human body, we can infer the possible region where the human body can lie in without penetrations. We refer to this region as the *free zone (FZ)*. Considering a person sitting on a chair with her/his legs under a table in front of the chair, the free zone is mainly the space between the chair and the table. With the free zone, we can significantly reduce the solution space for searching plausible poses of the invisible body parts. Second, inspired by the concept of shadow volume in computer graphics [Cro77], we consider the camera as a point light source, and construct a *truncated shadow volume (TSV)* to constrain the possible space for the visible body parts. The main idea is that the partially scanned body should fit and locate "behind" the seen point cloud in the direction of camera ray. In summary, in stead of estimating the human pose inside the whole space, we use above two strategies to make a confined region, within which the the 3D human pose is searched. By doing this, our results can avoid most penetrations between the body and the environment.

We design two methods based on above strategies. First, inspired by neural implicit fields [PFS*19, MON*19, CMPM20, KYZ*20, XBPM22], we use a neural network to estimate the free zone. Given a randomly sampled point in the whole space, the network can predict two field values: the body field value and scene field value. We construct the free zone by collecting those points whose body field value is below a certain threshold. Second, we compute the shadow volume of the body point cloud by shooting rays from the camera toward each point in the scanned body point cloud until they hit the environment. We further truncate the shadow rays from the scanned body point cloud to a certain maximum length limit and build the truncated shadow volume. We represent the truncated shadow volume discretely by uniformly sampled points along the truncated shadow rays. After obtaining the free zone and the truncated shadow volume, the next step is to match the body with them. We sample the human body by a differentiable interpolation algorithm, which produces points inside the body. Then we minimize the distances from the points in the body to points in the confined region. Our experiments show that such a volume matching method outperforms the surface matching method in terms of accuracy and robustness.

We also propose a more comprehensive metric to evaluate the penetration. Non-Collision (NC) is a traditional metric that measures the penetration between the body and the environment. It calculates the ratio of body vertices with positive Signed Distance Field (SDF) values. However, NC only considers body surface vertices and does not work well when part of the environment is inside the body. To address this limitation, we introduce a Volume Non-Collision (VNC) metric, which considers the points inside the body when calculating the similar ratio.

In summary, our contributions are as follows:

1. We propose two novel schemes, which consider the invisible and visible body part separately, to reduce the solution space for optimization-based pose reconstruction systems;
2. We design novel methods to apply above strategies by matching the body with the confined region as a volume;
3. We demonstrate that our system can reconstruct human poses with higher accuracy and less penetration compared to baseline methods.

2. Related Work

3D Human Reconstruction: 3D human reconstruction methods can be divided into optimization-based and regression-based methods. Optimization-based methods aim to reconstruct the 3D human body that matches with the RGB or RGB-D image by iteratively optimizing the parameters of the human body model [BKL*16, PCG*19, HCTB19, ZZB*21, LSS*22]. The key difference of these methods lie in the objective function, which typically consists of two parts: data terms and regularization terms [TZLW22]. Data terms are designed to align the human body with the input data, including RGB data and depth data. Regularization terms are used to constrain the parameters and prevent unrealistic poses and shapes. Bogo et al. [BKL*16] propose to iteratively fit the SMPL [LMR*15] human body model to the 2D keypoints detected by DeepCut [PIT*16]. Pavlakos et al. [PCG*19] follow a similar scheme but provide more detailed output for hands and face. However, these methods often suffer from visual artifacts, including scene penetration, feet sliding, and body leaning [TZLW22]. To

address these limitations, recent research has focused on leveraging scene information to constrain the pose and produce more plausible results. Hassan et al. [HCTB19] propose a human-scene penetration term and a contact term on top of SMPLify-X [PCG*19]. By considering the scene constraint, they achieve more realistic results with less penetration and necessary contact. Given partial observations, Zhang et al. [ZZB*21] introduce a motion smoothness prior to address jittering issues and employ a contact-aware motion infiller to infer plausible motions of occluded body parts. Compared with PROX-D [HCTB19], they produce improved results with smoother motions and more plausible body-scene interactions. In contrast to methods that treat the scene as a rigid object, Li et al. [LSS*22] jointly optimize the human body and the non-rigid deformation of the scene, leading to superior accuracy in reconstructing the 3D human body compared with other methods. Regression-based methods directly regress the parameters of the human body model in an end-to-end manner [KBJM18]. Kanazawa et al. [KBJM18] design a deep neural network to predict the parameters of SMPL human body model without requiring the 3D paired data. Kolotouros et al. [KPBD19] incorporate a post-process optimization module based on HMR [KBJM18] to improve the precision of the result. Our method is based on the optimization backbone, and it considers two new constraints: the free zone term and truncated shadow volume term to reduce the solution space.

Neural Implicit Field: The 3D model can be represented explicitly or implicitly. Recently, many works have been focused on using the implicit functions such as the DeepSDF [PFS*19], Occupancy Networks [MON*19], and UDF [CMPM20] to represent the 3D shape. There are also works using this method to represent the relationship between two objects [KYZ*20, XBPM22]. Karunratanakul et al. [KYZ*20] represent the hand and the grasped object using implicit field including the signed distances to the hand and the object. This representation can be used for grasp generation. Xie et al. [XBPM22] propose to extract the the body distance field, object distance field, object pose field, and body part field from an RGB image. These fields are used to optimize the parameters of the human body and the object, facilitating accurate and realistic reconstruction. Our method leverages a similar approach to model the relationship between the human body and the scene.

Interaction Representation: Various methods have been proposed to extract the interaction feature between two parts, such as a hand and a object or a human body and the surrounding environment [ZWK14, SHX*22]. Zhao et al. [ZWK14] propose to extract the Interaction Bisector Surface (IBS) between two objects using a geometry-based method, and use this feature for the classification and retrieval of 3D objects. She et al. [SHX*22] use IBS to represent the gripper-object interaction between gripper and object to solve the high-DOF reaching-and-grasping problem. The study of interaction feature between the human body and the surrounding scene has been explored in the field of 3D human reconstruction and generation [ZHN*20, ZZM*20, HGT*21]. Zhang et al. [ZHN*20] use the conditional Variational Auto-Encoder (cVAE) [SYL15] to predict semantically plausible 3D human body based on latent scene features. The generated human body are further refined by incorporating scene constraints to ensure plausible interactions. Zhang et al. [ZZM*20] model the proximal relationship between the human body and the scene using BPS [PLR19] feature. Hassan et al. [HGT*21] propose a body-centric human-scene interaction model that can be generalized to new scenes. These approaches contribute to the understanding and representation of human-scene interactions in the context of 3D human reconstruction and generation. Different from above methods that constrain the human pose in latent space, we explicitly reduce the solution space, within which the human pose is searched to align with the image and avoid penetrations.

3. Method

Our goal is to reconstruct the human body mesh from a monocular RGB-D image and the scene mesh. Our main idea is to reduce the solution space for body parts based on their visibility. In this section, we present details of our method.

3.1. Preliminaries

SMPL-X human body model: The SMPL-X [PCG*19] human body model is a differentiable function used to model the human body. It takes shape parameters $\beta \in \mathbb{R}^{12}$, pose parameters $\theta \in \mathbb{R}^{K\times 3}$, facial expression parameters ψ, and global translation $t \in \mathbb{R}^3$ as input. The output is a human body mesh $M_b = (V_b, F_b)$ composed of N_b vertices:

$$M(\beta,\theta,\psi,t) : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|\times|t|} \rightarrow \mathbb{R}^{N_b\times 3} \tag{1}$$

The pose parameters include poses for the body, hands, and jaw with axis-angle representation. K represents the total number of joints in the model, including 22 for the body, 30 for the hands (15 per hand), and 3 for the face. $J(\beta)$ is the 3D coordinates of each joint, which can be inferred from the vertices of the human body mesh using linear blend skinning. The parameters of the SMPL-X human body model can be optimized by adjusting the coordinates of the vertices or joints.

SMPLify-D and PROX-D: SMPLify-D and PROX-D [HCTB19] reconstruct the human body to align with the RGB-D data by optimizing the parameters of the SMPL-X human body model. The objective function is represented as follows:

$$E_{\text{SMPLify}-\text{D}} = E_J + \lambda_d E_d + \lambda_r E_r \tag{2}$$

E_J is a re-projection loss that aims to minimize the robust weighted distance between 2D joints estimated from the RGB image using OpenPose [CHS*21] and the 2D projections of the corresponding posed 3D joints of SMPL-X human body model. The depth term E_d minimizes the distances between the visible body vertices $V_b^v \in V_b$ and scanned body point cloud P_b which is extracted using the depth image and the body segmentation mask detected by DeepLab V3 [CPSA17] from the RGB image. The regularization term E_r is composed of multiple terms including pose prior term, shape prior term, self-penetration term, et al. Specifically, the self-penetration term [PCG*19] is used to avoid collisions between different body parts. λ_d and λ_r denote the weights for the depth term and regularization term respectively. PROX-D extends

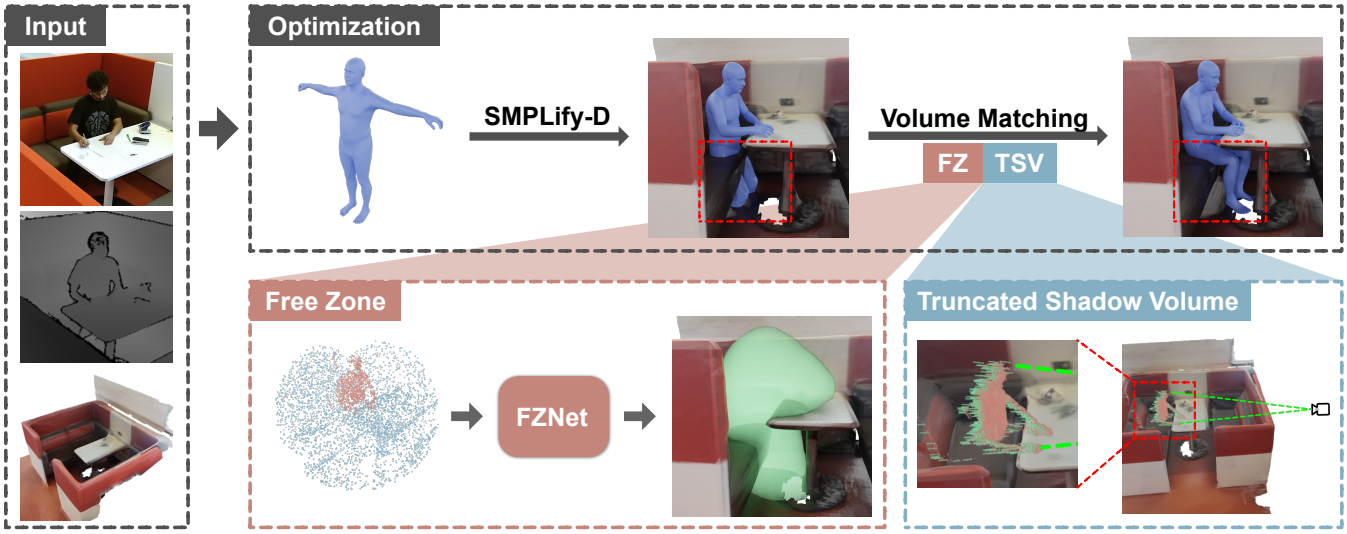

Figure 2: *The overview of our method. The input include a monocular RGB-D image and the scene mesh. In the first stage, we optimize the SMPL-X parameters from T-pose using SMPLify-D to get an initial result. In the second stage, we employ two strategies to reduce the solution space and use a volume matching algorithm to match the human body with the confined region. We design the free zone network (FZNet) to estimate the region where the human body can be positioned, which is used to constrain the invisible body parts; we calculate the truncated shadow volume behind the scanned body point cloud and use it to constrain the visible body parts.*

the SMPLify-D framework by adding a penetration term E_p and a contact term E_c to enforce scene constraint on the human body:

$$E_{\text{PROX}-\text{D}} = E_{\text{SMPLify}-\text{D}} + \lambda_p E_p + \lambda_c E_c \tag{3}$$

The penetration term E_p penalizes all penetrating vertices using the Signed Distance Field (SDF) of the scene mesh M_s. The contact term E_c encourages contact and proximity between the body and the scene by minimizing the distances from body vertices to scene vertices around contact areas. λ_p and λ_c denote the weights for the penetration term and contact term respectively.

3.2. Overview

Our optimization framework takes a monocular RGB-D image and the scene mesh as input. The output is the human body mesh. As depicted in Figure 2, our optimization framework consists of two stages. In the first stage, we use SMPLify-D to obtain an initial result. In the second stage, we employ two strategies to reduce the solution space and use the volume matching algorithm to match the human body with the confined region. The objective function of our method is defined as follows:

$$E = E_{\text{SMPLify}-\text{D}} + \lambda_{\text{fz}} E_{\text{fz}} + \lambda_{\text{tsv}} E_{\text{tsv}} + \lambda_c E_c \tag{4}$$

We extend SMPLify-D by introducing the free zone term E_{fz} and truncated shadow volume term E_{tsv} into the objective function. To prevent the issue of the body floating and encourage necessary contact, we also incorporate the contact term from PROX-D. λ_{fz} and λ_{tsv} denote the weights for the free zone term and truncated shadow volume respectively.

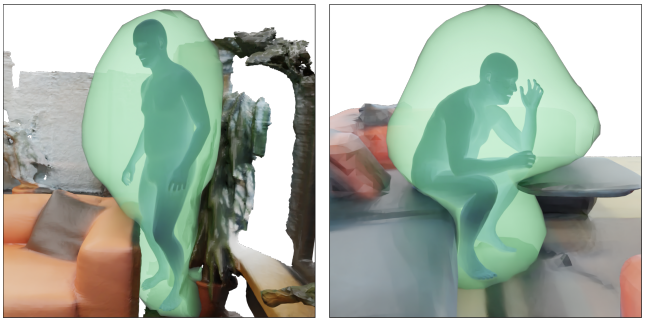

Figure 3: *Examples of the free zone and the corresponding reconstructed body.*

3.3. Free Zone

Current methods tries to solve the penetration problem by penalizing all penetrating body vertices using the SDF of the scene [HCTB19]. The effectiveness of this intuitive approach heavily relies on the accuracy and completeness of the scene. In reality, limitations in the scanning devices and the complexity of the scene can cause errors. Besides, this approach becomes even more ineffective in scenarios where the body part penetrates into the scene deeply or penetrates through thin objects. To solve these problems, we introduce a novel approach that uses a neural network to learn the

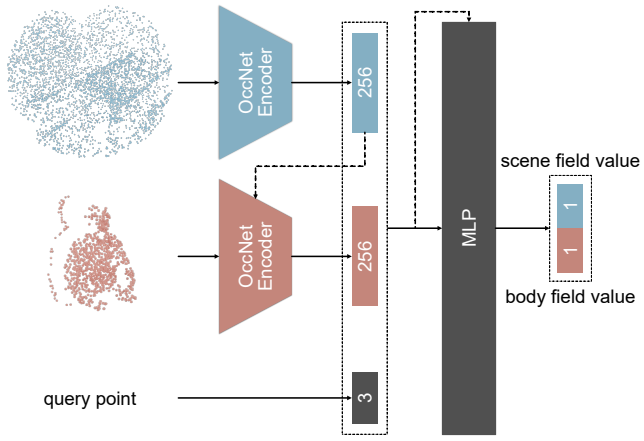

Figure 4: *The structure of the free zone network (FZNet). The input are the scene point cloud, the scanned body point cloud, and a query point. The output are the body field value and scene field value.*

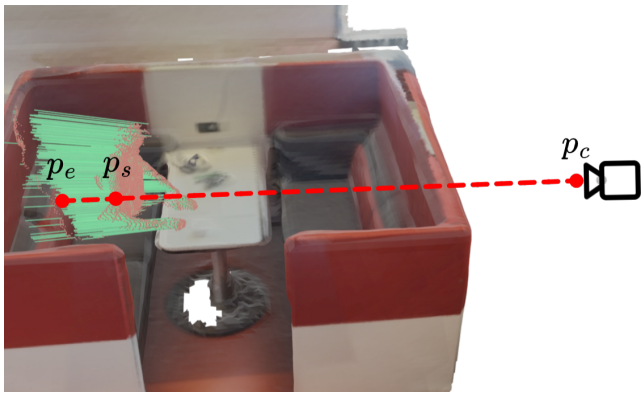

Figure 5: *The illustration of calculating the shadow volume points from the scanned body point cloud. Pink points denote the scanned body point cloud and green points denote the shadow volume points. The maximum length limit for truncation is set to infinity for better visualization.*

potential region where the human body can lie in without penetrating the environment. This region is referred to as the "free zone", and it serves as a confined region within which we search for plausible poses of the invisible body parts. In Figure 3, we show some examples of free zone and corresponding reconstructed body.

We propose to use an encoder-decoder-style network, the free zone network (FZNet), to estimate the free zone. The structure of FZNet is shown in Figure 4. The scanned body point cloud P_b and scene point cloud P_s are encoded separately using the OccNet encoder [MON*19]. The encoded scene point cloud feature is fed back to the body point cloud encoder to enforce condition. Then the scene feature, the body feature and the query point p are concatenated together and fed to a MLP decoder to predict the the field value for p in body field F_b and scene field F_s:

$$\text{FZNet}(P_b, P_s, p) = (F_b(p), F_s(p)) \tag{5}$$

To extract P_b, we first detect the body segmentation mask from the RGB image using DeepLab V3 [CPSA17]. Then, we obtain the whole point cloud, which includes both the human and the scene, from the depth image. Finally, we extract the body point cloud from the whole point cloud using the segmentation mask and obtain 1024 points using Farthest Point Sampling (FPS). P_s is sampled on the scene mesh with 4096 points. We train the free zone network by minimizing the L_1 distance between clamped prediction and ground truth distance:

$$\begin{aligned} L_{\text{FZNet}} = \sum_{p \in P_q} (&|\min(F_b(p), \delta) - \min(\text{GT}_b(p), \delta)| + \\ &|\min(F_s(p), \delta) - \min(\text{GT}_s(p), \delta)|) \end{aligned} \tag{6}$$

where P_q are the sampled query points. δ is the clamping distance and we set it as 0.1.

After getting the initial pose, we use the trained free zone network to get the free zone points. Assuming that the root joint of current human body is J_{root}. First, we sample scene point cloud within an unit sphere from scene mesh M_s centered at J_{root}. Next, we uniformly sample query points P_{grid} within a voxel grid with side length of 2 and resolution of 64 centered at J_{root}. For each sampled query point, we query its body field value. We then retain points with field value below a certain threshold, and consider these points as the free zone points used in the second stage of the optimization:

$$P_{\text{fz}} = \{p | F_b(p) < \mu, p \in P_{\text{grid}}\} \tag{7}$$

where μ is the threshold and we set it as 1e-3.

3.4. Truncated Shadow Volume

Existing methods may generate implausible results that appear in front of the scanned body point cloud. However, the visible body parts should be located behind the scanned body point cloud in the direction of the camera ray. We propose to calculate the truncated shadow volume behind the scanned body point cloud and use it to constrain the visible body parts.

We represent the shadow volume by shadow rays that are cast from the camera toward the scanned body. In Figure 5, we show the shadow rays, from which we can compute the truncated shadow rays. Let p_c represent the camera coordinate. The ray direction r is calculated as:

$$r = \frac{p_s - p_c}{\| p_s - p_c \|} \tag{8}$$

where $p_s \in P_b$ denotes the body point, which is also the start point

of the truncated shadow rays. We set a maximum length limit d_l to truncate the shadow rays, ensuring that the truncated shadow ray is not far from the scanned body. The distance from the first intersection point of the ray with the scene mesh to the start point is d_i. Then the end point p_e is computed by:

$$p_e = p_s + \min(d_l, d_i) \cdot r \tag{9}$$

By adding offsets to the start point p_s at a fixed interval, we can obtain a set of intermediate points that gradually move toward the end point p_e. These intermediate points form the truncated shadow volume points P_{tsv}. We empirically set the maximum length limit d_l as 0.1 and the interval as 0.01.

3.5. Differentiable Volume Matching

We propose to represent the human body as a volume composed of points located inside the body. To implement this, we need to obtain the internal points from body vertices while ensuring the process is differentiable.

As illustrated in Figure 6, we use an interpolation-based method to get the internal points from vertices. We use the SMPL-X human body model at T-pose as a template mesh to get the interpolation vertex pairs. We first divide the body vertices into front vertices and back vertices by calculating the visibility using a virtual camera in front of the body. Then we sample front vertices using FPS to make the vertices uniformly distributed. For each front vertex v_f, we cast a ray towards the back vertices. We can get the nearest back vertex v_b from the intersection point of the ray with the human body mesh. v_f and v_b form an interpolation vertex pair. By repeating above steps for all selected front vertices, we can obtain a set of interpolation vertex pairs. Then we perform linear interpolation and get 6 points for each pair, resulting in the internal points P_{int}. Assuming the body is locally convex, P_{int} will not extend beyond the body boundary and can remain the body pose and shape.

Free zone can be seen as a superset of the body and truncated shadow volume can be seen as a subset of the body. We match the human body with free zone points and truncated shadow volume points separately. For the free zone points, we employ the following loss:

$$E_{\text{fz}} = \sum_{p_i \in P_{\text{int}}} \rho_{\text{fz}}(\min_{p_j \in P_{\text{fz}}} \| p_i - p_j \|) \tag{10}$$

where ρ_{fz} denotes a robust Geman-McClure error function [GEM87] for down weighting the points in P_{fz} that are far from P_{int}, so that the human body will not become too fat. For the truncated shadow volume points, we employ the following loss:

$$E_{\text{tsv}} = \sum_{p_i \in P_{\text{tsv}}} \rho_{\text{tsv}}(\min_{p_j \in P_{\text{int}}} \| p_i - p_j \|) \tag{11}$$

where ρ_{tsv} denotes a robust Geman-McClure error function for down weighting points in P_{int} that are far from P_{tsv}, so that the human body will not become too thin.

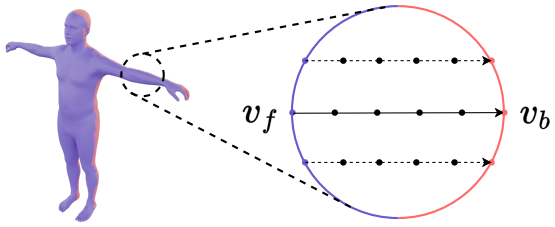

Figure 6: *The process of getting the internal points from body vertices. Purple denotes the front and red denotes the back.*

4. Experiments

4.1. Datasets

We conduct experiments on the PROX dataset [HCTB19]. The PROX dataset is divided into two parts: a qualitative set and a quantitative set. The qualitative set consists of monocular videos of 20 human subjects interacting with 12 indoor scenes. The dataset contains 100K RGB-D frames recorded at 30 fps, along with the scene mesh. The pseudo-ground-truth SMPL-X [PCG*19] parameters are fitted using PROX-D [HCTB19]. The quantitative set consists of 180 static RGB-D frames, with one human subject wearing markers and interacting with a living room containing daily furniture. The ground-truth SMPL-X parameters are fitted using MoSh++ [MGT*19]. We use the scene data from POSA [HGT*21], which aligns the scene data of the PROX dataset for easy processing.

4.2. Experiment Details

Dataset Split: We randomly split the qualitative set into training and testing sets with a ratio of 4:1. The training set is used to train the free zone network. We exclude the data that has high penetrations and extract frames every second. The testing set is used to evaluate our method. We also conduct experiments on the quantitative set.

Free Zone Network Training: We sample 20K query points every training sample. Among these points, 95% are located near the surface and 5% are uniformly distributed within the bounding box. To generate points near the surface, we first sample surface points on both the body and the scene. Then we introduce perturbations to each surface point along three axes. These perturbations are generated by applying zero-mean Gaussian noise with variances of 0.02 and 0.002, resulting in two query points every surface point. For each generated point, we calculate its nearest distance to the body points and scene points. This allows us to determine the point's proximity to the body and scene, providing valuable information for training the free zone network. To enhance the generalization of the free zone network, we apply two data augmentation techniques, including random rotation along the z-axis and online query points generation. By employing these data augmentation techniques, we aim to create a diverse training set that enables the free zone network to generalize on different visible ratios and scene types. We

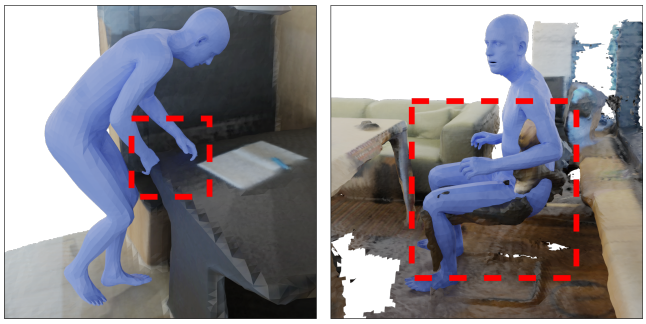

Figure 7: *Examples to illustrate that Non-Collision can not evaluate the penetration correctly. The penetration area is highlighted using red dashed box.*

train the free zone network using the Adam optimizer [KB14] with a learning rate of 1e-4. We employ a step learning rate schedule with a decay rate of 0.5 after 100 epochs. The models are trained for 200 epochs on a single 3090Ti GPU.

Optimization: We use the scene as the world coordinate and both the free zone and truncated shadow volume are defined in this coordinate. Before matching with the confined region, we transform the human body from the camera coordinate to the world coordinate. The weights for each term in the objective function are set empirically. Specifically, we set the weight of the free free zone term to 5e2 and the weight of the truncated shadow volume term to 2e2. We use the same optimizer with SMPLify-X [PCG*19].

4.3. Results

Figure 8 presents a gallery of our results. We can see that our method can produce accurate and plausible poses on different scene types and visible ratios. When the human body is sitting on a chair and only the upper body is visible (row 1, 2), the free zone can guide the human body to be in a sitting pose, no matter the human is sitting back or facing to the camera. When the human has a lot of contact with the scene (row 3), such as when the human is lying on a sofa, our method can produce plausible results with little penetration and necessary contact. In cases where the scenes are more complex (row 4), such as when the human is standing between a sofa and a pot of plants, our free zone network can still identify the correct region. Even when the human is not captured by the depth camera (row 5), the free zone network can still estimate the free zone correctly using only the scene information. Additionally, when the human is in an irregular pose (row 6), our method can still produce results that match with the image.

4.4. Comparison and Evaluation

We compare our method with SMPLify-D and PROX-D [HCTB19] using a series of evaluation metrics on both PROX quantitative and qualitative sets.

Evaluation Metrics: We evaluate the performance using a set of metrics, which can be categorized into accuracy metrics and plausibility metrics. We assess the 3D reconstruction accuracy using Joint Position Error (JPE) and Vertex-to-Vertex Error (V2V). JPE measures the mean per joint position error and V2V represents the mean error between corresponding vertices of the reconstructed and the ground truth mesh. We also apply Procrustes alignment to the meshes and calculate the aligned JPE (p.JPE) and aligned V2V (p.V2V). To evaluate the alignment accuracy with the depth data, we use a Partial Matching (PM) metric, which measures the mean distance between each body point and its closest vertex on the reconstructed mesh. The PM score is lower when the matching is more accurate.

To evaluate the plausibility, we employ the Non-Collision (NC) metric introduced in PSI [ZHN*20]. This metric is calculated as the ratio of body vertices with a positive SDF value. The NC score is lower when the penetration is more severe. However, we find this metric does not work well in scenarios where the environment is inside the body. For example, in Figure 7, the NC scores of these two examples are both 0.93. However, the penetrations are completely different. In the left example, only part of the hands penetrates the table. In the right example, the chair penetrates into the body completely. This highlights a potential discrepancy in evaluating penetration using the NC. To address this problem, we propose a modified version called Volume Non-Collision (VNC), which calculates the ratio of the internal points with a positive SDF value. The VNC score is lower when the penetration is more severe. By considering the points inside the body, this metric provides a more comprehensive evaluation of penetration. In Figure 7, the VNC scores of these two examples are 0.99 and 0.79 respectively, suggesting a more accurate evaluation of penetration compared with the NC.

Quantitative Results: The comparison results on the quantitative set are listed in Table 1. Our method achieves the lowest error in all metrics, although there is only one scene in the quantitative set and the interactions are simple.

Table 1: *Results on PROX quantitative set.*

	JPE ↓	V2V ↓	p.JPE ↓	p.V2V ↓
SMPLify-D	73.80	76.81	45.61	44.57
PROX-D	69.46	72.70	42.43	42.20
Ours	**66.74**	**70.04**	**41.86**	**41.46**

Qualitative Results: The comparison results on the qualitative set are listed in Table 2. Our method achieves the best performance in all metrics.

Table 2: *Results on PROX qualitative set.*

	E_p	E_c	E_{fz}	E_{tsv}	NC ↑	VNC ↑	PM/1e-3 ↓
SMPLify-D					95.6%	95.2%	4.42
PROX-D	✓	✓			96.8%	97.3%	4.53
Ours		✓	✓	✓	**97.1%**	**97.9%**	**4.10**

In Figure 9, we visually compare our method with other methods on the PROX qualitative set. In the first 4 rows, where the human is partially occluded by the scene or by themselves, both SMPLify-D and PROX-D produce results where some parts penetrate the

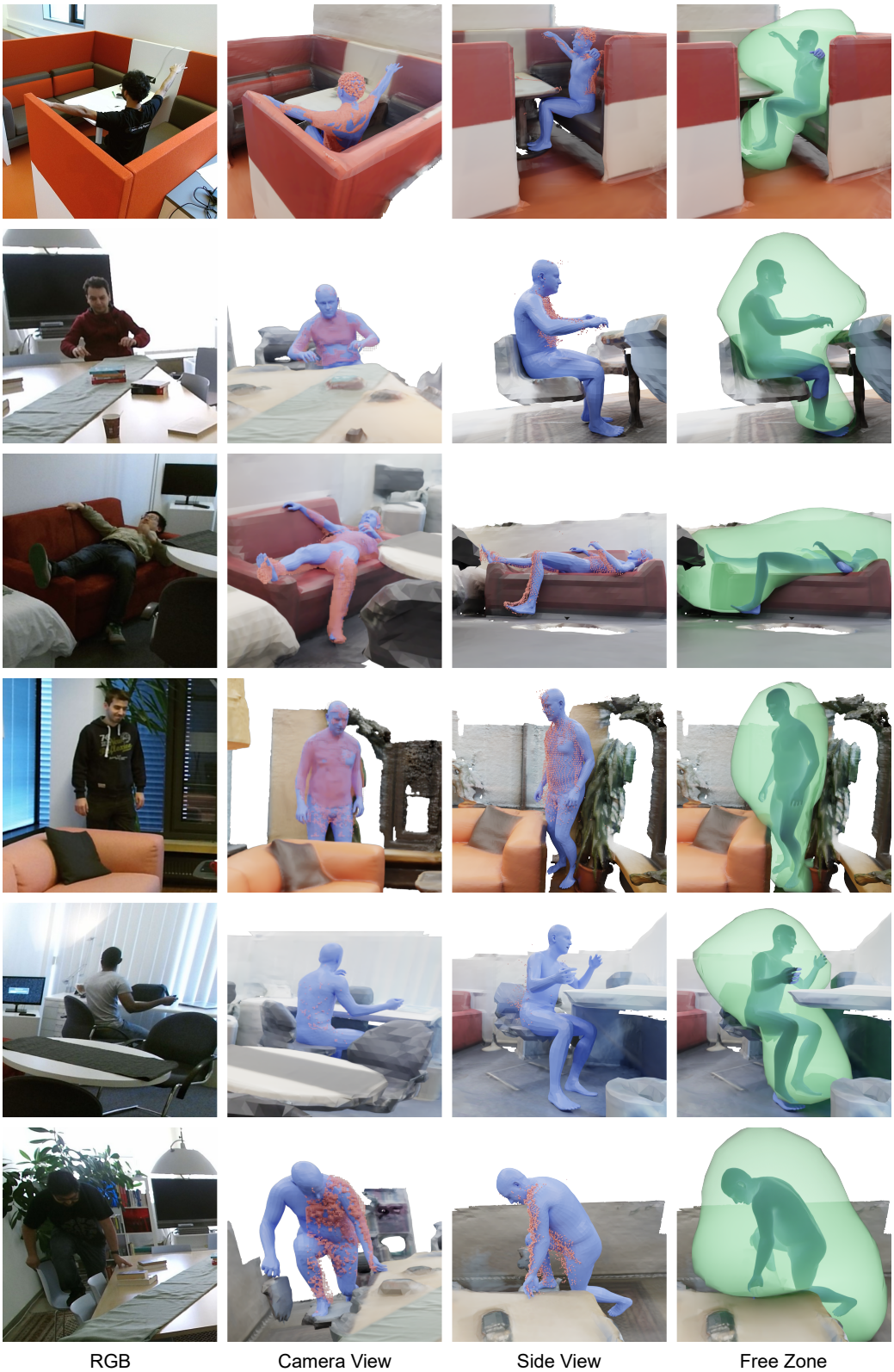

Figure 8: *Gallery of our results. We show examples from different scenes and poses. Pink points denote the scanned body point cloud.*

Table 3: *Evaluation based on different scene types.*

	Sitting Booth			Chair			Sofa			Bed		
	NC ↑	VNC ↑	PM/1e-3 ↓	NC ↑	VNC ↑	PM/1e-3 ↓	NC ↑	VNC ↑	PM/1e-3 ↓	NC ↑	VNC ↑	PM/1e-3 ↓
SMPLify-D	93.8%	93.1%	8.97	96.9%	96.5%	3.17	96.7%	96.2%	3.75	92.6%	92.3%	5.14
PROX-D	95.8%	96.4%	9.20	97.4%	97.8%	3.38	97.5%	97.9%	4.14	95.5%	96.4%	5.41
Ours	**96.1%**	**97.3%**	**8.44**	**97.6%**	**98.2%**	**3.04**	**97.8%**	**98.4%**	**3.56**	**96.5%**	**97.3%**	**4.72**

Table 4: *Evaluation based on different visible ratios.*

	0% – 25%			25% – 50%			50% – 75%			75% – 100%		
	NC ↑	VNC ↑	PM/1e-3 ↓	NC ↑	VNC ↑	PM/1e-3 ↓	NC ↑	VNC ↑	PM/1e-3 ↓	NC ↑	VNC ↑	PM/1e-3 ↓
SMPLify-D	94.9%	91.1%	12.40	94.3%	93.8%	5.71	96.6%	96.7%	2.85	**99.8%**	**99.8%**	1.16
PROX-D	97.1%	96.9%	14.62	95.9%	96.4%	5.97	97.4%	98.1%	2.69	99.3%	99.7%	**1.12**
Ours	**97.6%**	**98.0%**	**12.00**	**96.4%**	**97.3%**	**5.35**	**97.7%**	**98.3%**	**2.56**	99.2%	99.7%	1.13

scene. However, our method can infer the correct pose of the invisible body parts and avoid penetrations. When some body parts penetrate deeply into objects, such as a leg penetrating into a sofa or a hand penetrating into a wall (row 1, 2), it is hard for current methods to pull the body out of the object completely. However, our method uses the free zone to guide the body away from the object, effectively reducing the penetration. Our method can also handle cases where some body parts penetrate through thin objects like a table (row 3, 4), preventing such penetrations from occurring. In the last 2 rows, where different body parts overlap with each other, our results exhibit better alignment with the scanned body point cloud compared with other methods thanks to the constraint of the truncated shadow volume.

We further analyse of the qualitative results by scene types and visible ratios. Results based on different scene types are shown in Table 3. We can see that our method has a more significant improvement on scenes with sitting booths or beds. In scenes with sitting booths, the legs may be occluded by the table or chair. Current methods often reconstructs the wrong pose with the leg penetrating the chair and struggles to pull the leg out. Our method is more capable of handling such occlusions with the aid of the free zone term. For scenes with beds, the geometry of the bed can serve as a strong clue to generate accurate free zone and truncated shadow volume, leading to more accurate and plausible reconstruction. Results based on different visible ratios are shown in Table 4. The visible ratio is calculated by comparing the scanned body point cloud with the reconstructed results of PROX-D. We can see that our method has a more significant improvement when the visible ratio is lower, which demonstrates the effectiveness of our method in situations with serious occlusions.

4.5. Ablation Study

We conduct the ablation study on the PROX qualitative set and the results are shown in Table 5. We consider the following ablation versions:

- Ours (w/o FZ): we test how our method performs when the free zone term is removed.
- Ours (w/o TSV): we test how our method performs when the truncated shadow volume term is removed.
- Ours (w/o VM): we replace the volume matching with surface matching.

We can see that the final version which includes all the proposed components achieves the best overall performance.

Table 5: *Ablation study on PROX qualitative set.*

	E_{fz}	E_{tsv}	VM	NC ↑	VNC ↑	PM/1e-3 ↓
Ours (w/o FZ)		✓	✓	95.1%	95.4%	4.18
Ours (w/o TSV)	✓		✓	96.6%	96.4%	4.47
Ours (w/o VM)	✓	✓		96.2%	96.3%	4.25
Ours	✓	✓	✓	**97.1%**	**97.9%**	**4.10**

Without the free zone term, the reconstructed results have more penetrations with the scene, and both the NC score and VNC score decrease. Figure 10 presents two representative examples to demonstrate this effect. In the first example, when our method is applied without the free zone term, the reconstructed result shows that the left leg penetrates into the sitting booth. In the second example, we can observe a similar scenario where the right hand penetrates into the wall. By using the free zone term to constrain the invisible body parts, we can effectively reduce the penetration issue, ensuring a more plausible reconstruction.

Without the truncated shadow volume term, the reconstructed results do not match with the scanned body point cloud well and the PM score decreases. Figure 11 presents two representative examples to demonstrate this effect. In the first example, when our method is applied without the truncated shadow volume term, the reconstructed result displays a mismatch between the left hand and the scanned body point cloud, leading to an unnatural pose. In the second example, although the pose is natural, there is a clear mismatch between the body and the RGB image because the right hand is positioned incorrectly. These examples highlight the important role played by the truncated shadow volume term. By incorporating the truncated shadow volume term, there is a better alignment with the scanned body point cloud, resulting in a visually coherent reconstruction.

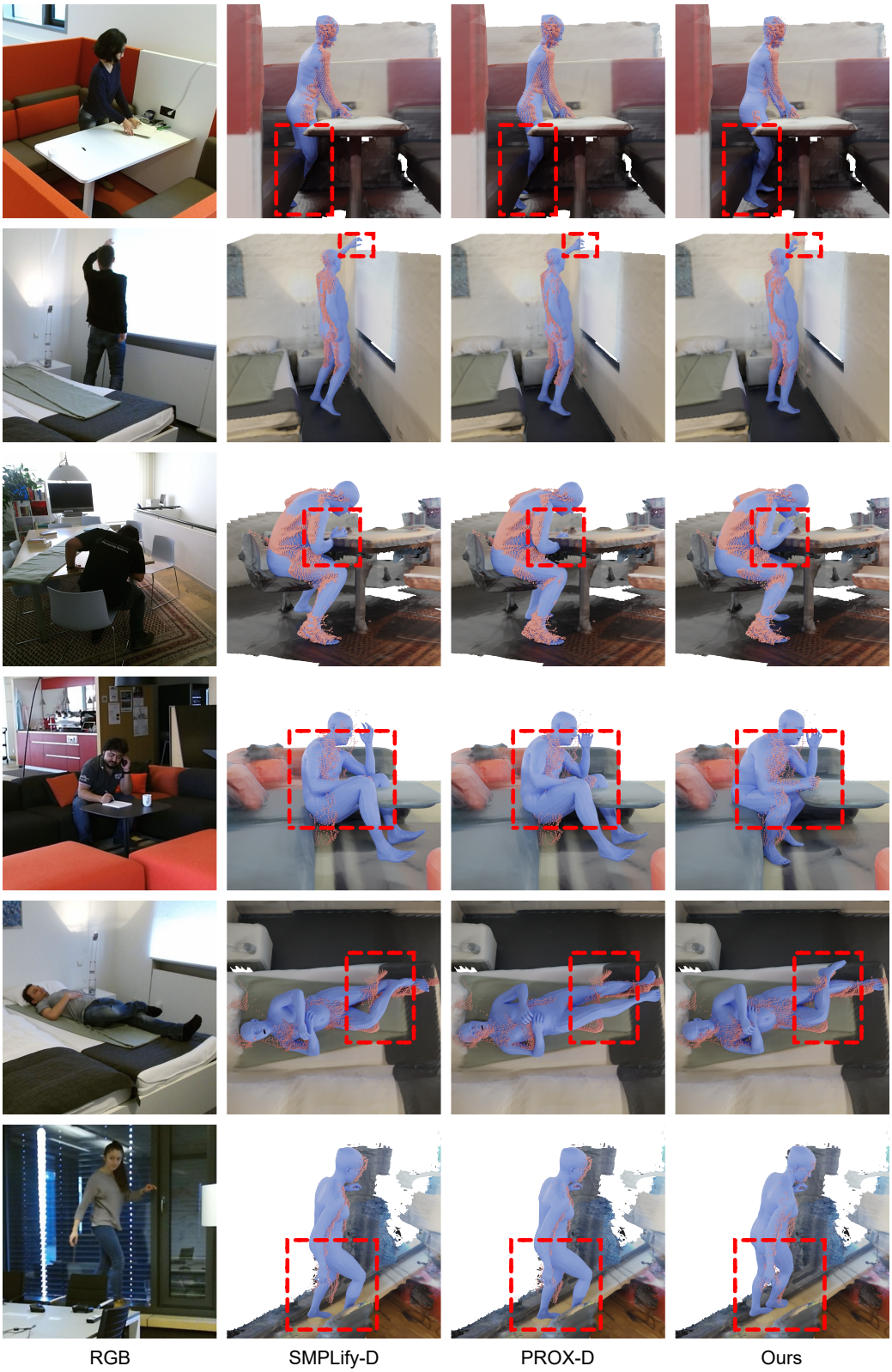

Figure 9: *Comparison of the reconstructed results by our method with those of SMPLify-D and PROX-D. Pink points denote the scanned body point cloud. The difference between results is highlighted using red dashed box.*

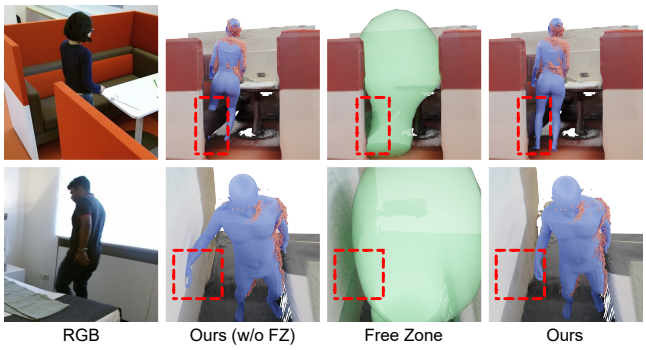

Figure 10: *Ablation study: compare our method with the version that does not use free zone term. Pink points denote the scanned body point cloud. The difference between results is highlighted using red dashed box.*

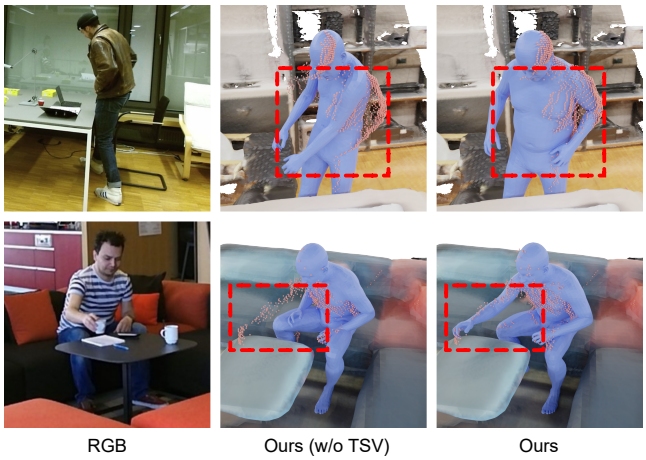

Figure 11: *Ablation study: compare our method with the version that does not use truncated shadow volume term. Pink points denote the scanned body point cloud. The difference between results is highlighted using red dashed box.*

When the volume matching is replaced with surface matching, the effectiveness of the free zone term and truncated shadow volume term both decrease, and all metrics have a slight decline. Figure 12 presents two representative examples to demonstrate this effect. In the first example, when volume matching is replaced with surface matching, the reconstructed result does not match with the scanned body point cloud correctly. In the second example, the human body does not match with the free zone correctly and the left leg penetrates into the sofa. After applying the volume matching algorithm, the reconstructed result is more accurate and plausible. These examples demonstrate the importance of volume matching algorithm. By matching with the confined region using internal points, the free zone and truncated shadow volume can have a more significant effect.

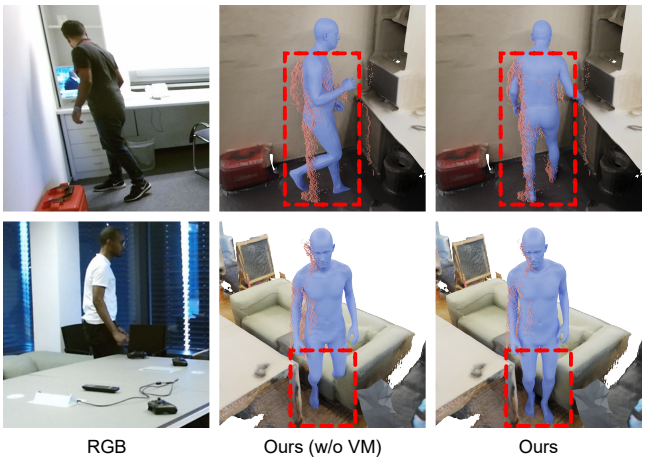

Figure 12: *Ablation study: compare our method with the version that uses surface matching instead of volume matching. Pink points denote the scanned body point cloud. The difference between results is highlighted using red dashed box.*

5. Conclusion and Discussions

We present a novel method for 3D human body reconstruction from a monocular RGB-D image and the scene mesh. We introduce two schemes to explicitly reduce the solution space based on scene information and prior knowledge. We show that the free zone term can reduce the penetrations and the truncated shadow volume term can increase the matching accuracy. Additionally, a novel volume matching algorithm is proposed to match the human body with the confined region, which yields better performance than surface matching. We also introduce a more comprehensive metric, Volume Non-Collision (VNC), to evaluate the penetration by considering the entire body as a volume. Extensive experiments demonstrate that the proposed method produces more accurate and plausible results compared to other methods, especially in situations with close interactions and serious occlusions.

Limitations and future work: Our method is deterministic and can only reconstruct one result every time. However, there exists multiple possible results due to occlusion and diversity of pose. In the future, we will explore how to reconstruct multiple possible results which are diverse and plausible. Furthermore, our method assumes a rigid scene, neglecting the deformations that may occur during human-scene interactions in real-world scenarios. To enhance the realism of the reconstructed results, we consider incorporating scene deformations into our approach in future work.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62072366), National Key R&D Program of China (2022YFB3303200), Special Key Projects of Guiding Technological Innovation in Shaanxi Province: 2021QFY01-03 and Key Plan of New Technology & New Business in Tang Dou Hospital: XJSXYW2021001.

References

[BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P., ROMERO J., BLACK M. J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14* (2016), Springer, pp. 561–578. 2

[CHS*21] CAO Z., HIDALGO G., SIMON T., WEI S.-E., SHEIKH Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 1 (2021), 172–186. doi:10.1109/TPAMI.2019.2929257. 3

[CMPM20] CHIBANE J., MIR A., PONS-MOLL G.: Neural unsigned distance fields for implicit function learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2020), NIPS'20, Curran Associates Inc. 2, 3

[CPMAMN22] CORONA E., PONS-MOLL G., ALENYÀ G., MORENO-NOGUER F.: Learned vertex descent: A new direction for 3d human model fitting. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II* (Berlin, Heidelberg, 2022), Springer-Verlag, p. 146–165. URL: https://doi.org/10.1007/978-3-031-20086-1_9, doi:10.1007/978-3-031-20086-1_9. 2

[CPSA17] CHEN L.-C., PAPANDREOU G., SCHROFF F., ADAM H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017). 3, 5

[Cro77] CROW F. C.: Shadow algorithms for computer graphics. *SIGGRAPH Comput. Graph. 11*, 2 (jul 1977), 242–248. URL: https://doi.org/10.1145/965141.563901, doi:10.1145/965141.563901. 2

[GEM87] GEMAN S.: Statistical methods for tomographic image restoration. *Bull. Internat. Statist. Inst. 52* (1987), 5–21. 6

[HCTB19] HASSAN M., CHOUTAS V., TZIONAS D., BLACK M.: Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 2282–2292. doi:10.1109/ICCV.2019.00237. 1, 2, 3, 4, 6, 7

[HGT*21] HASSAN M., GHOSH P., TESCH J., TZIONAS D., BLACK M. J.: Populating 3d scenes by learning human-scene interaction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 14703–14713. doi:10.1109/CVPR46437.2021.01447. 3, 6

[JNV21] JOO H., NEVEROVA N., VEDALDI A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)* (2021), pp. 42–52. doi:10.1109/3DV53792.2021.00015. 2

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 7

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7122–7131. doi:10.1109/CVPR.2018.00744. 2, 3

[KPBD19] KOLOTOUROS N., PAVLAKOS G., BLACK M., DANIILIDIS K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 2252–2261. doi:10.1109/ICCV.2019.00234. 2, 3

[KYZ*20] KARUNRATANAKUL K., YANG J., ZHANG Y., BLACK M. J., MUANDET K., TANG S.: Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)* (2020), pp. 333–344. doi:10.1109/3DV50981.2020.00043. 2, 3

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM Trans. Graph. 34*, 6 (nov 2015). URL: https://doi.org/10.1145/2816795.2818013, doi:10.1145/2816795.2818013. 2

[LSS*22] LI Z., SHIMADA S., SCHIELE B., THEOBALT C., GOLYANIK V.: Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *2022 International Conference on 3D Vision (3DV)* (2022), pp. 1–11. doi:10.1109/3DV57658.2022.00013. 2, 3

[LWL21] LIN K., WANG L., LIU Z.: End-to-end human pose and mesh reconstruction with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 1954–1963. doi:10.1109/CVPR46437.2021.00199. 2

[MGT*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M.: Amass: Archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 5441–5450. doi:10.1109/ICCV.2019.00554. 6

[MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3d reconstruction in function space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4455–4465. doi:10.1109/CVPR.2019.00459. 2, 3, 5

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10967–10977. doi:10.1109/CVPR.2019.01123. 2, 3, 6, 7

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 165–174. doi:10.1109/CVPR.2019.00025. 2, 3

[PIT*16] PISHCHULIN L., INSAFUTDINOV E., TANG S., ANDRES B., ANDRILUKA M., GEHLER P., SCHIELE B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4929–4937. doi:10.1109/CVPR.2016.533. 2

[PLR19] PROKUDIN S., LASSNER C., ROMERO J.: Efficient learning on point clouds with basis point sets. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019), pp. 3072–3081. doi:10.1109/ICCVW.2019.00370. 3

[SCH*16] SAVVA M., CHANG A. X., HANRAHAN P., FISHER M., NIESSNER M.: Pigraphs: Learning interaction snapshots from observations. *ACM Trans. Graph. 35*, 4 (jul 2016). URL: https://doi.org/10.1145/2897824.2925867, doi:10.1145/2897824.2925867. 2

[SHX*22] SHE Q., HU R., XU J., LIU M., XU K., HUANG H.: Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction. *ACM Trans. Graph. 41*, 4 (jul 2022). URL: https://doi.org/10.1145/3528223.3530091, doi:10.1145/3528223.3530091. 3

[SYL15] SOHN K., YAN X., LEE H.: Learning structured output representation using deep conditional generative models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA, USA, 2015), NIPS'15, MIT Press, p. 3483–3491. 3

[TZLW22] TIAN Y., ZHANG H., LIU Y., WANG L.: Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923* (2022). 2

[WCR*19] WANG Z., CHEN L., RATHORE S., SHIN D., FOWLKES C.: Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718* (2019). 2

[XBPM22] XIE X., BHATNAGAR B. L., PONS-MOLL G.: Chore: Contact, human and object reconstruction from a single rgb image. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II* (Berlin, Heidelberg, 2022), Springer-Verlag, p. 125–145. URL: https://doi.org/10.1007/978-3-031-20086-1_8, doi:10.1007/978-3-031-20086-1_8. 2, 3

[ZHN*20] ZHANG Y., HASSAN M., NEUMANN H., BLACK M. J., TANG S.: Generating 3d people in scenes without people. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6193–6203. doi:10.1109/CVPR42600.2020.00623. 3, 7

[ZMZ*22] ZHANG S., MA Q., ZHANG Y., QIAN Z., KWON T., POLLEFEYS M., BOGO F., TANG S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI* (2022), Springer, pp. 180–200. 2

[ZWK14] ZHAO X., WANG H., KOMURA T.: Indexing 3d scenes using the interaction bisector surface. *ACM Trans. Graph. 33*, 3 (jun 2014). URL: https://doi.org/10.1145/2574860, doi:10.1145/2574860. 3

[ZZB*21] ZHANG S., ZHANG Y., BOGO F., POLLEFEYS M., TANG S.: Learning motion priors for 4d human body capture in 3d scenes. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 11323–11333. doi:10.1109/ICCV48922.2021.01115. 2, 3

[ZZM*20] ZHANG S., ZHANG Y., MA Q., BLACK M. J., TANG S.: Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)* (2020), pp. 642–651. doi:10.1109/3DV50981.2020.00074. 3