


# Multi-Level Implicit Function for Detailed Human Reconstruction by Relaxing SMPL Constraints

Xikai Ma<sup>1</sup>, Jieyu Zhao<sup>2</sup>, Yiqing Teng<sup>1</sup>, Li Yao<sup>†1,3</sup> 

<sup>1</sup>School of Computer Science and Engineering, Southeast University, China

<sup>2</sup>Southeast University - Monash University Joint Graduate School (SuZhou), China

<sup>3</sup>Key Lab. of Computer Network and Information Integration (Southeast University), Ministry of Education, China

## Abstract

*Aiming at enhancing the rationality and robustness of the results of single-view image-based human reconstruction and acquiring richer surface details, we propose a multi-level reconstruction framework based on implicit functions. This framework first utilizes the predicted SMPL model (Skinned Multi-Person Linear Model) as a prior to further predict consistent 2.5D sketches (depth map and normal map), and then obtains a coarse reconstruction result through an Implicit Function fitting network (IF-Net). Subsequently, with a pixel-aligned feature extraction module and a fine IF-Net, the strong constraints imposed by SMPL are relaxed to add more surface details to the reconstruction result and remove noise. Finally, to address the trade-off between surface details and rationality under complex poses, we propose a novel fusion repair algorithm that reuses existing information. This algorithm compensates for the missing parts of the fine reconstruction results with the coarse reconstruction results, leading to a robust, rational, and richly detailed reconstruction. The final experiments prove the effectiveness of our method and demonstrate that it achieves the richest surface details while ensuring rationality. The project website can be found at <https://github.com/MXKKK/2.5D-MLIF>.*

## CCS Concepts

• **Computing methodologies** → **Computer vision; Shape modeling;**

## 1. Introduction

Image-based human 3D reconstruction is a key problem in computer vision and graphics research, with widespread applications in VR/AR content creation [CPW\*18], entertainment, video editing and enhancement [HTCH15, FP09], holographic [OERF\*16], virtual dressing [PMPHB17], and more. In the past, to obtain human 3D models, expensive scanning equipment was required, as well as the expenditure of time and labor for scanning and various post-processing algorithms to fill scanning gaps, improve mesh quality, and so on. With technological advancements and the rise of deep learning, image-based human 3D reconstruction has gradually become a research hotspot. Its objective is to provide single or multiple human body images from different angles as input and obtain 3D mesh reconstructions of the human body for downstream applications. Based on the number of input images, it can be divided into multi-view human 3D reconstruction and single-view human 3D reconstruction.

In the field of human 3D reconstruction, single-view human 3D

reconstruction is more important because a photo containing human is generally taken from one direction, and the quality requirements for reconstruction are relatively higher. This includes accurately reconstructing facial features as well as the folds of clothing. Consequently, deep learning methods have gradually become the mainstream approach for human 3D reconstruction.

In recent years, numerous methods for single-view human body reconstruction have emerged. Methods based on parametric human body models first appeared [BKL\*16, APMTM19]. These parametric models condense the human body into dozens of parameters [LMR\*15], significantly reducing the learning difficulty for neural networks and yielding reasonable reconstruction results. In the past two years, there have been numerous advancements and extensions in explicit methods, leading to more accurate parameterized human body model estimation [CPMAMN22] and clothed human body reconstruction [MNSL22]. However, such methods can only reconstruct predefined human bodies and severely lack details, such as clothing folds. Voxel-based methods [ZYW\*19, VCR\*18] have also gained popularity in recent years due to their compatibility with deep learning techniques. The biggest issue with these methods is the selection of voxel resolution. Choosing a lower resolution results in a lack of detail in the reconstruction, while choos-

<sup>†</sup> Significant Science And Technology Project of Nanjing under Grant No. 202209003

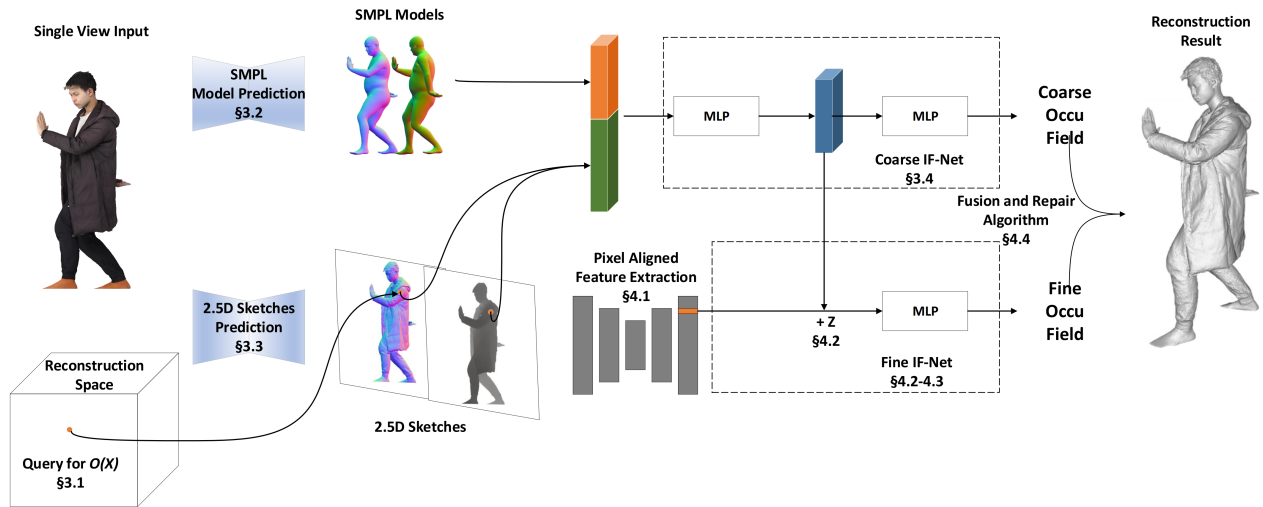


Figure 1: The pipeline of our method

ing a higher resolution leads to memory overuse and reconstruction failure.

There are also point cloud-based human body reconstruction methods [GFM\*19]. These methods seem to underperform due to the non-uniqueness of point cloud representations (i.e., the same triangular mesh can be represented by different point clouds). Implicit function methods have achieved great success. These methods first extract features from the input image and then use a neural network to fit an implicit function. Given any point in space as input, the occupancy value of that point is obtained, indicating whether the point is inside or outside the reconstructed 3D object. This approach solves the resolution limitation problem of voxel-based methods, with the reconstruction resolution determined by the input query points. Since the obtained implicit function is continuous, points can be requested arbitrarily in space and their implicit function values calculated, allowing surface reconstruction using the MarchingCubes algorithm [LC87]. The performance of such methods depends on the accuracy of the implicit function obtained, with representative methods including PIFu [SHN\*19], PIFuHD [SSSJ20], and TetraTSDF [OHT\*20].

Implicit function-based methods can reconstruct detailed human body surfaces, but their biggest issue is under-constraint, meaning the input image does not provide enough information and there is no suitable parametric model. This leads to a lack of detail in invisible parts and even produces unreasonable reconstruction results. Currently, fully utilizing the limited information from image inputs, providing sufficient constraints for reconstruction results, and achieving reasonable, high-quality reconstructions remain pressing problems to be solved.

At present, single-view human 3D reconstruction based on images still mainly relies on implicit function-based methods. The current research trend is to provide prior constraints for implicit function fitting using various techniques, such as using predicted SMPL models as priors [ZYLD21, XYTB22], predicting front and

back normal maps as priors [SSSJ20], predicting voxel reconstructions as priors [HCJS20], and so on. However, some of these methods cannot reconstruct complex human body poses, some still lack detail, and some produce cluttered and unreasonable surface reconstructions. Many existing methods use the SMPL model as a prior constraint, aiming to obtain reasonable reconstructions for various complex posture inputs. However, the overly strong constraint of the SMPL model on the reconstruction results limits the presentation of surface details. How to enhance the presentation of surface details while ensuring the rationality of the reconstruction results is still an unresolved issue.

In response to these issues, we make the following contributions:

1. In order to enhance surface details while ensuring the rationality of the reconstruction, we propose a novel multi-level reconstruction framework. We first obtain a reasonable and robust reconstruction through the coarse implicit function fitting network guided by SMPL models. Then, by using a fine implicit function fitting network based on relaxing SMPL model priors, we reduce the influence of the SMPL model on the results while ensuring the rationality of the reconstruction, thereby achieving a reconstruction with richer surface details.
2. To address the issue of possible missing hands and feet in reconstruction results under complex poses with high-precision reconstruction, we propose a coarse-fine occupancy field fusion and repair algorithm. This algorithm detects missing parts and fills them in with smoothed coarse occupancy fields, further enhancing the reasonableness of the reconstruction.

The pipeline of our method is shown in Figure 1. First, a coarse implicit function fitting network (coarse IF-Net) guided by the SMPL model is proposed, using the predicted SMPL model and 2.5D sketches as input to obtain a reasonable and robust initial reconstruction for various input poses. This part is elaborated in Section 3. Furthermore, in order to add surface details, a fine IF-Net aimed at relaxing the constraints of the SMPL model is designed.



It guides the reconstruction results towards more detailed directions by extracting pixel-aligned features from the 2.5D sketches. All SMPL model-related features are replaced, retaining only the global features obtained from the coarse IF-Net, allowing the reconstruction results to break through the limitations of the SMPL model, and achieve more detailed reconstruction. This part is elaborated in Section 4.

In Section 5, experimental results show that the reconstruction method proposed in this paper is advanced and effective.

## 2. Related Work

**Single-View Image Based 3D Reconstruction:** Single-view image based 3D reconstruction is an ill-posed problem due to its ambiguous nature. Unlike multi-view image-based 3D reconstruction, single-view image-based 3D reconstruction often requires strong priors. In recent years, deep learning-based approaches have demonstrated promising and exciting results. Wang et al. [WZL\*18] employed an initial ellipsoid mesh to deform into a final mesh. Several methods [WSK\*15, CXG\*16, TZEM17] attempted to recover a voxel representation from a single image. To mitigate the memory footprint caused by high-resolution voxels, space partitioning techniques (e.g., octree) were employed [ROUG17, HTM19, TDB17]. In order to further achieve infinite resolution, signed distance fields (SDF) [PFS\*19, XWC\*19, BTFB21, MON\*19] were utilized, and the surface was extracted using the marching cubes algorithm [LC87]. In general, single-view image-based 3D reconstruction can be summarized as encoding input images into a specific feature space and then decoding them into various 3D representations (such as point clouds [YSR\*20], voxels, SDF, etc.).

**Image Based Human Reconstruction:** Image-based human reconstruction can be divided into template methods and template-free methods. Template methods attempt to regress the parameters of a predefined template. SMPL [LMR\*15] and STAR [OBB20] are commonly used human body templates, and several studies [BKL\*16] have attempted to fit a human body template into a single RGB image. Template-free methods have achieved significant success in recent years and are mainly divided into voxel-based methods, point cloud-based methods, and implicit function-based methods. Voxel-based methods typically obtain a voxel space of the human body through an encoder-decoder structure (e.g., U-Net [RFB15], 3D U-Net [ÇAL\*16]) and then carry out a series of post-processing steps (such as minimizing the reprojection loss) to obtain the 3D reconstruction of the human body [VCR\*18, ZYW\*19]. Gabeur et al. [GFM\*19] infer the front and back depth maps and perform point cloud sampling, while Jinka et al. [JCSN20] obtain more depth information by simulating light propagation. Both of these methods perform Poisson reconstruction [KBH06] to reconstruct human bodies from point clouds. Saito et al. [SHN\*19] introduced the first method that uses an implicit function for human body reconstruction, and many implicit function-based methods have followed [HXL\*20, SSSJ20]. Recently, several methods have combined two of the four representations mentioned above. [ZYLD21, XYTB22] combined the SMPL model with implicit function-based methods, while He et al. [HCJS20] combined voxel-

based methods with implicit functions. Hong et al. [HZJ\*21] utilized multi-view stereo to achieve better results.

With the significant success of NeRF [MST\*21] in recent years, there have also been NeRF-Based methods [SYZR21, SBR22] that produce high-quality reconstruction results through neural rendering.

## 3. Coarse Implicit Function Fitting Guided by the SMPL Model

The process of coarse implicit function fitting guided by the SMPL model follows steps similar to those in the ICON [XYTB22] method, with a key distinction: during ICON’s normal prediction stage, our method adds a depth map prediction that is consistent with the normal map. This addition enhances the rationality of the reconstruction results. Moreover, the prediction of the depth map plays a crucial role in relaxing the SMPL constraints. Overall, this part first predicts the SMPL model, then predicts the normal map based on the SMPL model, and next conducts a cooperative optimization of the SMPL model parameters and the normal map. Ultimately, a consistent depth map prediction is achieved based on the normal map prediction. In the subsequent implicit function fitting phase, we use the SMPL features and the 2.5D sketches features as inputs to the implicit function, thereby obtaining a coarse occupancy field.

### 3.1. Implicit Functions and Occupancy Fields

First, we define the occupancy field as in Equation (1). Given a query point  $X$  in the reconstruction space, the occupancy field value is 0 when  $X$  is outside the 3D object, and 1 in all other cases:

$$O(X) = \begin{cases} 0, & \text{if } X \text{ outside the surface} \\ 1, & \text{else} \end{cases} \quad (1)$$

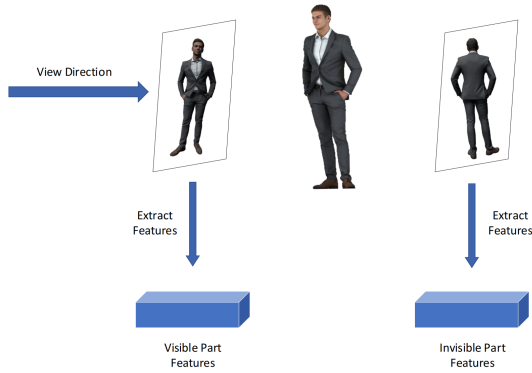
We want to fit an implicit function  $f$  to obtain this occupancy field  $O$ , as shown in Equation (2):

$$f(S(F_{vis}, F_{invis}, X), F_{query}) = O(X) \quad (2)$$

where  $F_{query}$  is the query point features which are independent of the visibility, and  $F_{vis}$  and  $F_{invis}$  are the visible and invisible part features, respectively. As shown in Figure 2, the visible part features are obtained from the features extracted from the image in the view direction, while the invisible part features are extracted from the image predicted in the opposite direction of the view.  $F_{query}$  represents certain features inherent in the query point itself. During the coarse implicit function fitting phase where the SMPL model is used as a prior constraint,  $F_{query}$  is denoted as  $F_{query}^{coarse}$  and takes the form of Equation (3):

$$F_{query}^{coarse}(X) = [F_s(X), F_n^b(X)] \quad (3)$$

Here,  $F_s(X)$  is the signed distance from a query point  $X$  to the closest point  $X^b$  on the SMPL model, and  $F_n^b$  is the barycentric surface normal of  $X^b$ . As the SMPL constraints are relaxed in the fine implicit function fitting stage,  $F_{query}$  will take a different form and will discuss later in Section 4.2.



**Figure 2:** The visible and invisible part features

$S$  in Equation (2) is a selection function related to how to deal with visible and invisible part features. In the coarse IF-Net, we use SMPL visibility for selection, as in Equation (4).

$$S^{coarse}(F_{vis}, F_{invis}, X) = \begin{cases} F_{vis}, & \text{if } X^b \text{ is visible} \\ F_{invis}, & \text{else} \end{cases} \quad (4)$$

Taking  $F_{vis}$  as an example, it is calculated as in Equation (5), and  $F_{invis}$  has a similar form:

$$F_{vis} = \Phi(x, D, N) \quad (5)$$

where  $x$  is the projection of  $X$  in the pixel space, usually taking the first two components of the  $X$  coordinate.  $D$  and  $N$  are the previously predicted normal and depth maps. In the coarse implicit function fitting phase,  $\Phi$ , denoted as  $\Phi^{coarse}$ , simply takes out the corresponding values from the depth and normal maps, as in Equation (6). While in the subsequent fine implicit function fitting phase,  $\Phi^{fine}$  represents the pixel-aligned feature extractor, as discussed in Section 4.1 and Equation (9).

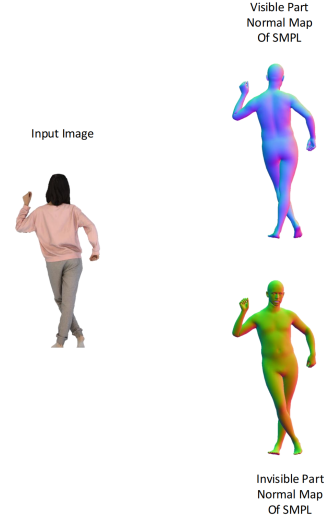
$$\Phi^{coarse} = [D(x), N(x)] \quad (6)$$

### 3.2. SMPL Model Prediction

The SMPL model [LMR\*15] is a parameterized human skin triangle mesh that can describe a human body with 10 shape parameters and 24 pose parameters. It provides a strong prior constraint for the reconstruction results. Just like many methods [ZYL21, XYTB22], we use SMPL model to guide our coarse reconstruction result.

The prediction of the SMPL model uses the PyMaf method [ZTZ\*21], which can predict a reasonable SMPL model from a single image and align its triangular mesh with the original image. This method extracts features at different scales and gradually refines the prediction results from coarse to fine granularity, aligning the mesh with the image.

In our implementation, the SMPL model prediction network directly uses the pre-trained network provided by PyMaf. Once the SMPL model is obtained, we render it into visible and invisible part normal maps, as well as their signed distance fields for later use. The input image and its corresponding SMPL model normal map are shown in Figure 3.



**Figure 3:** SMPL model prediction

### 3.3. 2.5D Sketches Prediction

2.5D sketches refer to 2D images that can express some 3D information. Some typical 2.5D sketches are shown in Figure 4, including the following types:

- **Depth map:** a single-channel grayscale image, with each pixel storing the depth of the 3D object in the current camera space (i.e., the Z-axis coordinate). The darker the color, the closer it is to the camera; the lighter the color, the farther it is from the camera.
- **Normal map:** a three-channel image, with each pixel storing the local normal direction of the 3D object at that pixel point. The surface details of the 3D object can be seen from the normal map.
- **Silhouette map:** a mask with only two values, 0 and 1, indicating whether a pixel point belongs to the 3D object. It plays an essential role when the background is complex or there are multiple objects.

[GFM\*19] demonstrated the effectiveness of predicting depth maps, [SSSJ20] proved that predicting normal maps could enhance the details of reconstruction results. And silhouette maps are indispensable for complex backgrounds.

At this stage, we extract the silhouette map of the human body and further predict the visible part normal and depth maps, and the invisible part normal and depth maps. These images are collectively referred to as 2.5D sketches for later use. The silhouette map can be predicted using image segmentation techniques, and we focus on predicting depth and normal maps. The 2.5D sketches prediction process proposed in this paper is shown in Figure 5. The 2.5D



Figure 4: 2.5D sketches

sketches prediction network has a similar structure to the generator of the Pix2PixHD [WLZ\*18]. Since invisible part prediction is required, we concatenate the SMPL normal map obtained in the SMPL feature extraction stage with the input image and input it into the prediction network to obtain visible and invisible part normal predictions. Subsequently, to ensure the consistency between SMPL model predictions and normal predictions, SMPL refinement [XYTB22] is used. SMPL refinement optimizes the SMPL model parameters as variables, using the rendered SMPL model normal maps and silhouette maps as loss terms. This process helps bring the predicted SMPL model closer to the input images, thereby further enhancing the quality of the reconstruction. Following that, to ensure the consistency between depth and normal predictions, the input image and predicted normals are concatenated and input into the depth prediction network, resulting in visible and invisible part depth predictions.

The 2.5D sketches prediction network has a loss function as shown in Equation (7):

$$E = E_{l1}(N_{GT}, N_P) + E_{vgg}(N_{GT}, N_P) \quad (7)$$

It consists of an L1 loss and a VGG loss, where  $N_{GT}$  represents the actual depth or normal map, and  $N_P$  represents the depth or normal map predicted by the 2.5D sketches prediction network. In our implementation, the depth prediction network uses only the L1 loss. The training of the 2.5D sketches prediction network is conducted separately. In the training set, we have real normal maps, depth maps, and accurate SMPL models, thus making the training of the 2.5D sketches prediction network relatively straightforward.

### 3.4. Coarse Implicit Function Fitting

The IF-Net is composed of a Multilayer Perceptron (MLP) that generates an occupancy value output for each query point. That



Figure 5: 2.5D sketches prediction process

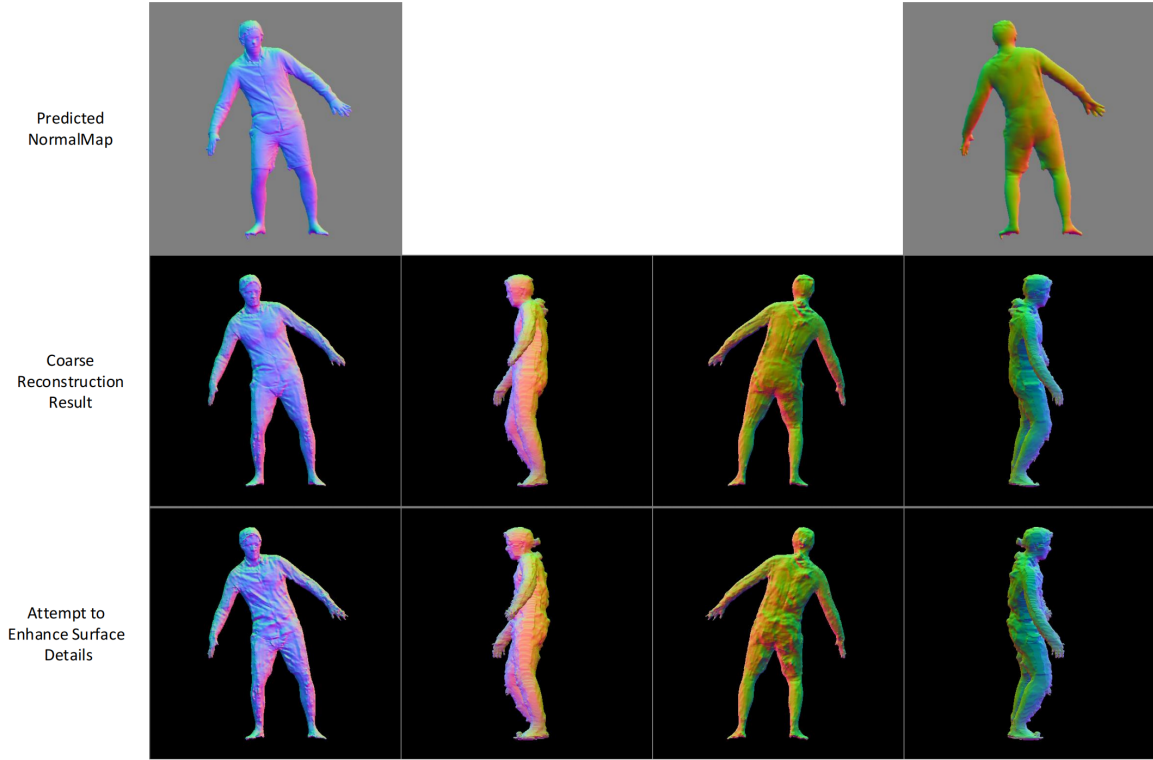
is, the function  $f$  in Equation (2) is approximated using MLP. The MLP has a loss function as shown in Equation (8):

$$L = \frac{1}{n} \sum_{i=1}^n (f(X_i) - f^*(X_i))^2 \quad (8)$$

where  $f(X_i)$  is the actual occupancy value of the query point, and  $f^*(X_i)$  is the predicted occupancy value of the query point after passing through the implicit function fitting network.

## 4. Fine Implicit Function Fitting by Relaxing the Prior Constraints Imposed by the SMPL Model

As shown in the first and second rows of Figure 6, the coarse reconstruction results still retain some of the original features of the SMPL model, and there is a discrepancy in the surface details with the predicted normal map. We aim to achieve a reconstruction with richer surface details. We attempted to enhance the surface details of the reconstruction result using a multi-level structure, but the result is as shown in the third row of Figure 6. The reconstruction seems to be agonizingly choosing between the SMPL model and the normal map, resulting in chaotic outcomes. Thus, a method is needed to relax the overly strong constraints of the SMPL model



**Figure 6:** SMPL prior constraints limit surface details

to achieve a reconstruction that is rich in surface details without losing robustness.

This chapter introduces the multi-level structure we used, using pixel-aligned features and query point depth to replace the SMPL model priors, thereby relaxing the constraints of the SMPL model. In Section 5.6, we also elaborate on the trade-off between surface details and rationality that we discovered during the experiments.

#### 4.1. Using Pixel Aligned Feature to Emphasize 2.5D Sketches

Pixel-aligned feature was first proposed by [SHN\*19]. It represents a feature space where a corresponding feature vector can be obtained for each point in the input image space, thereby contributing to the prediction of the final implicit function values. In order to bring the reconstruction results closer to the 2.5D sketches with rich details, we use a pixel-aligned feature extractor to separately extract features from the visible and invisible parts of the 2.5D sketches. We use a stacked hourglass network as the pixel-aligned feature extractor [SHN\*19]. During the reconstruction process, the  $\Phi$  in Equation (5) becomes the pixel-aligned feature extractor. See Equation (9). Given the projection  $x$  of the query point  $X$ , and the 2.5D sketches  $D$  and  $N$ , the pixel-aligned feature of the query point can be obtained.  $\theta$  represents the parameters of the pixel-aligned feature extractor. The obtained pixel-aligned feature can be used for subsequent fine reconstruction. Note that features from visible

( $F_{vis}$ ) and invisible parts ( $F_{invis}$ ) are extracted separately. This is achieved through two independent pixel-aligned feature extractors.

$$\Phi^{fine} = \theta(x, D, N) \quad (9)$$

#### 4.2. Using Query Point Depth Instead of SMPL Model Prior

Recalling Equation (3),  $F_{query}^{coarse}$  is a series of information obtained by projecting the query point onto the closest face of the SMPL model, which carries strong priors of the SMPL model. To relax the priors of the SMPL model in the fine implicit function fitting stage, its form must be changed. As shown in Equation (10), we replace it with the depth value of the query point (since the  $(x, y)$  coordinates of the query point are already reflected in the projected point  $x$  of the pixel-aligned feature). Moreover, in order to further reduce the prior nature of the SMPL features, the feature selection using SMPL visibility is also removed in the fine IF-Net, and the visible and invisible parts are concatenated, as shown in Equation (11).

$$F_{query}^{fine}(X(x, y, z)) = z \quad (10)$$

$$S^{fine}(F_{vis}, F_{invis}, X) = [F_{vis}, F_{invis}] \quad (11)$$

### 4.3. Coarse IF-Net as Global Feature Extractor

However, the robustness provided by the SMPL model as a prior for the reconstruction of various complex posture input images is also necessary. We cannot completely disregard the SMPL priors. Therefore, we extract the intermediate layer of the Coarse IF-Net as the global feature of the query point, retaining some, but not strong, SMPL priors. This maintains the rationality and robustness of the reconstruction results. Thus, Equation (10) further changes to:

$$F_{query}^{fine}(X(x,y,z)) = [z, f_{Coarse}^4(X)] \quad (12)$$

where  $f_{Coarse}^4(X)$  delegate to extract the fourth layer of coarse IF-Net.

### 4.4. Fusion and Repair of Coarse and Fine Occupancy Fields

The strategy mentioned above also cause some problems, as shown in Figure 8. When the pose of the input image is too complex, especially when the pose is not covered in the training set, the reconstruction result may miss parts of the hands and feet. We speculate that this is why the P2S indicator in the comparison experiment in Section 5.4.1 is slightly higher for our method. This problem can be solved naturally by using a larger-scale and higher-quality training set. However, we noticed that the coarse reconstruction result, which uses the SMPL feature as input, can obtain a more complete reconstruction but lacks details and has more noise. Therefore, we propose a fusion and repair algorithm for coarse and fine occupancy fields. The algorithm process is shown in Algorithm 1. By detecting missing blocks in the fine occupancy field on the depth axis of the occupancy field (rather than finding the noisy part of

---

#### Algorithm 1: The Fusion and Repair Algorithm

---

**Input:** FineOccu, CoarseOccu

**Output:** FineOccu

CoarseOccu = smooth(CoarseOccu);

```

for  $i, j$  in Width(CoarseOccu), Height(CoarseOccu) do
  start = -1;
  fineConut = 0;
  for  $k$  in Depth(CoarseOccu) do
    if Entering Coarse Surface then
      start =  $k$ ;
      fineCount = 0;
    end
    if InsideFineOccu( $i, j, k$ ) then
      fineCount++;
    end
    if Exiting Coarse Surface then
      coarseCount =  $k$  - start;
      if  $fineCount / coarseCount < threshold$  then
        merge(CoarseOccu, FineOccu,  $i, j, start, k$ );
      end
    end
  end
end

```

---

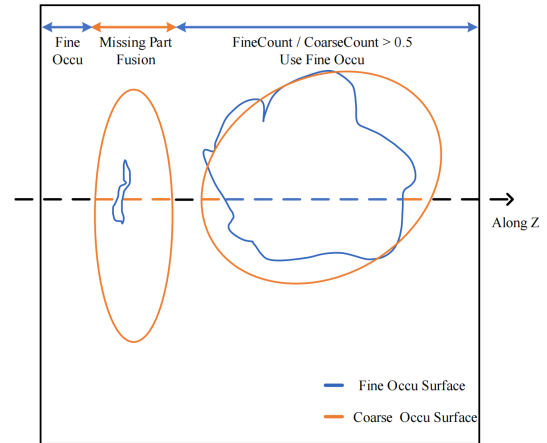


Figure 7: Fusion and repair algorithm visualization

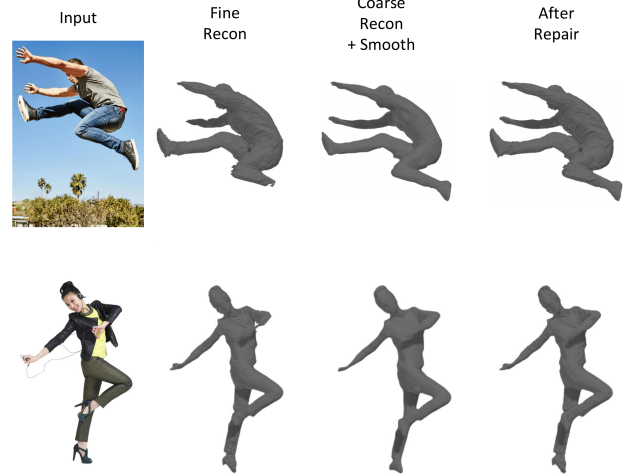


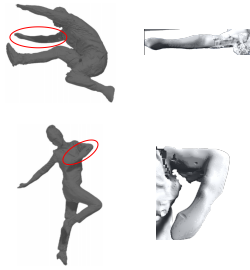
Figure 8: Fusion and repair of coarse and fine occupancy fields

the coarse occupancy field), we replace them with the smoothed coarse occupancy field to fill in the missing parts of the fine reconstruction result and achieve a more reasonable reconstruction. The merge function in Algorithm 1 blends the coarse occupancy field with the fine occupancy field smoothly.

In Figure 7, the visualization of the algorithm can be observed. First, we find the segments inside the coarse surface (indicated by the orange dashed lines) and study the proportion of points on this segment that are located within the fine surface. On the left side (the orange double-headed arrow), the proportion is relatively small ( $< 0.5$ ), which is determined to be the missing part. In this case, we use a fusion of the coarse occupancy field and the fine occupancy field. On the right side (the blue double-headed arrow), the fine occupancy field is preserved.

The effect of the fusion and repair method is shown in Figure 8 and Figure 9 provides a close-up view of the fusion part. It can





**Figure 9:** Zoomed-in details of the merge part

be seen that our algorithm effectively fills in the missing parts and minimizes the introduction of additional noise.

#### 4.5. Training Strategy

Since our method is composed of multiple modules, the training steps are also divided into several parts, specifically as follows:

1. Train the SMPL model prediction network (or use a pre-trained model, mentioned in Section 3.2), and train the 2.5D sketch prediction network (mentioned in Section 3.3) using the loss function as in Equation (7).
2. Train coarse IF-Net (mentioned in Section 3.4) using the loss function as in Equation (8).
3. Train fine IF-Net and pixel aligned feature extractor (mentioned in Section 4.1) together using the loss function as in Equation (8).

### 5. Experiments

#### 5.1. Datasets

We utilize THuman2.0 [YZG\*21] as the training set and CAPE [MYR\*20] as the test set. The data quality of THuman2.0 is superior, whereas the CAPE dataset exhibits more complex input poses and contains a certain level of noise, making it extremely suitable for testing.

#### 5.2. Implementation Details

The 2.5D sketches prediction network is trained separately. It consists of four downsampling modules, four ResNetBlocks, and four upsampling modules.

The coarse IF-Net is a MLP which has a dimension of (11, 256, 512, 256, 128, 1), and layers 3, 4, and 5 use residual connections. Its input is consists of 7-dimensional SMPL features and 4-dimensional 2.5D sketches features selected by the SMPL model visibility. The pixel-aligned feature extraction network uses the stacked hourglass network described in Section 4.1, consisting of two stacked hourglass networks. For each query point, visible and invisible parts can extract 128-dimensional pixel-aligned features, resulting in 256-dimensional pixel-aligned features. The fourth-layer features of the coarse IF-Net are then extracted, resulting in 256-dimensional global features. These are concatenated with the 256-dimensional pixel-aligned features and 1-dimensional query

point depth to form a 513-dimensional vector, which is the input to the fine IF-Net. The fine IF-Net has a dimension of (513, 512, 1024, 512, 256, 1). The third, fourth, and fifth layers use residual connections. During training, the RMSprop optimizer is used, the learning rate is set to 1e-4, and no weight decay is set. Similar to [SHN\*19], each group of training data uses 8000 query points, including points uniformly sampled from the reconstruction space and points obtained from importance sampling from the GroundTruth mesh surface. For training, the standard deviation for coarse IF-Net is set to 0.5, while for fine reconstruction network it is set to 0.3.

When training the coarse IF-Net, we downsample the 2.5D sketches to a size of 512 \* 512. During the training of the fine reconstruction network, we use a 512 \* 512 window to randomly crop the 2.5D sketches and then sample and train within the image range.

All experiments were conducted using an RTX 3090 graphics card with 24GB of VRAM. According to our observations, both the training and testing phases require a minimum of 15GB of VRAM.

#### 5.3. Quantitative Analysis Metrics

The reconstruction accuracy is evaluated using three quantitative analysis metrics. To assess the global accuracy and rationality of the reconstruction, the Chamfer Distance (CD) and Point to Surface Distance (P2S) between the reconstructed mesh and the real mesh are used as quantitative analysis indicators. The lower these two indicators, the closer the reconstructed mesh is to the real mesh, and the more reasonable the reconstruction. In addition, to evaluate the level of detail of the reconstructed surface, the normal loss (NI2) is introduced as a quantitative analysis metric, which involves rendering normal maps of the real mesh and the reconstructed mesh from four directions and then calculating the L2 loss. The lower the normal loss, the more detailed the reconstructed surface.

#### 5.4. Comparative experiment

In this section, we compare our method with state-of-the-art methods in terms of quantitative and qualitative aspects, demonstrating the superiority and effectiveness of our approach.

##### 5.4.1. Quantitative Analysis

Table 1 quantitatively analyzes the performance of different methods on the CAPE dataset, comparing our method with several commonly used single-view human body reconstruction methods. It can be seen that our method is significantly better than the current methods in terms of CD and NI2, while slightly worse than the ICON method in terms of P2S. As analyzed in Section 4.4, this may be due to the missing limbs at the end when the input pose is very complex, as is shown in Figure 10. However, the better CD indicates that our method is robust and reasonable overall. After introducing the fusion and repair algorithm proposed in Section 4.4, all indicators, especially P2S, have been further improved, achieving state-of-the-art performance. Therefore, our method greatly improves surface details while ensuring reasonable reconstruction results. This is also evident in the qualitative analysis of the reconstruction quality in the following sections. For the ClothWild, being

a parameterized model-based reconstruction approach, it achieves excellent P2S metrics. However, its surface severely lacks details that correspond to the input images, resulting in a poor NI2 metric. Further qualitative comparison is presented in Figure 11.



Figure 10: Almost the most severe condition of missing limbs

Table 1: Quantitative Comparison with Other Methods

Methods	CD↓	P2S↓	NI2↓
PIFu [SHN*19]	2.03	1.58	0.117
PaMIR [ZYLD21]	1.68	1.44	0.119
ICON [XYTB22]	1.29	1.15	0.087
ClothWild [MNSL22]	1.21	<b>0.86</b>	0.135
LVD [CPMAMN22]	4.72	5.32	0.236
Ours	<u>1.18</u>	1.26	<u>0.079</u>
Ours Wi/Repair	<b>1.14</b>	<u>1.12</u>	<b>0.078</b>

#### 5.4.2. Qualitative Analysis

Figure 11 presents a comparison between our method and the reconstruction results obtained by non-IF-Based methods. It is evident that our approach captures rich surface details, while also faithfully preserving input image features, such as bare feet, hair, and clothing wrinkles that align with the images.

Figure 12 presents a comparison of the reconstruction results between our method and other IF-Based methods. It can be seen that our method demonstrates robustness for more complex poses, producing reasonable results for various poses. Compared to the ICON method, our method preserves rich details while greatly reducing noise, resulting in clean and tidy outputs that can be used directly without post-processing.

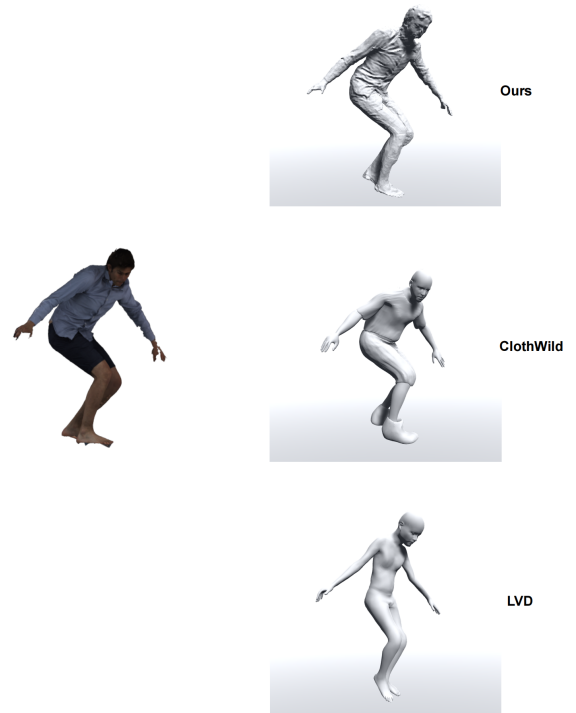


Figure 11: Comparison with non-IF-Based methods

#### 5.5. Ablation Study

In this section, we prove the effectiveness of our designed multi-layer IF-Net through quantitative and qualitative ablation studies. There are five settings for the ablation studies: (1) Applying our designed multi-level implicit function fitting network, denoted as: Ours; (2) Using only the coarse implicit function fitting network, denoted as: Coarse Only; (3) Using only the fine reconstruction network, denoted as: Fine Only; (4) Using only the coarse implicit function fitting network and adding the pixel-aligned feature extractor before the coarse implicit function fitting network, denoted as: Coarse Only Wi/ PF; (5) Using the multi-level implicit function fitting network and adding the pixel-aligned feature extractor before the coarse implicit function fitting network, denoted as: Ours Wi/CoarsePF. All settings are trained on the THuman2.0 dataset and tested on the CAPE dataset, yielding the quantitative metrics shown in Table 2. Note that all the settings did not use the fusion and repair algorithm.

Table 2: Quantitative Ablation Study

Settings	CD↓	P2S↓	NI2↓
Ours	<b>1.18</b>	1.26	<b>0.079</b>
Coarse Only	1.29	<b>1.15</b>	0.087
Fine Only	1.93	1.68	0.139
Coarse Only Wi/ PF	1.21	1.16	0.083
Ours Wi/CoarsePF	1.35	1.47	0.083

It can be seen that, except for the P2S metric analyzed before,



Figure 12: Comparison with IF-Based methods

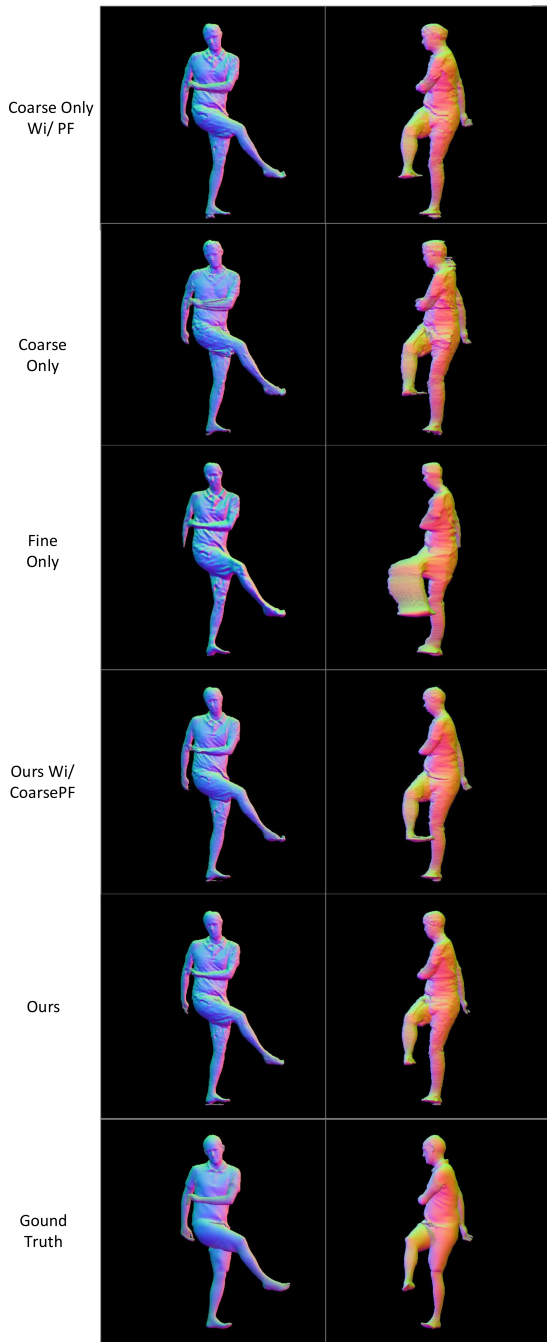


Figure 13: Qualitative ablation study

our method performs significantly better than other settings on the CD and NI2 metrics, indicating that our multi-resolution network structure is effective. The CD metric is better than other settings, and the P2S metric is not far behind other settings, indicating that our multi-resolution network structure effectively preserves the robustness and rationality of the single-layer network. The NI2 nor-



Figure 14: Qualitative ablation study

mal loss is significantly lower than other settings, indicating that our multi-resolution network structure can increase surface detail.

Figure 13 show the comparison of the reconstruction results obtained from the five settings and the visualization of the Ground Truth mesh. It can be seen that the Coarse Only setting results in significant noise, and its surface appears chaotic; the Coarse Only Wi/PF setting produces a more reasonable reconstruction, but its surface lacks detail; the Fine Only setting focuses more on surface details but lacks rationality; Ours Wi/CoarsePF, although increasing the number of parameters, seems to lack rationality; Ours, while

maintaining the rationality of the reconstruction results, obtains the richest surface details. To further demonstrate the surface details of the Ours setting, Figure 14 shows the comparison of surface details with the Coarse Only Wi/PF setting.

### 5.6. Trade-off Between Surface Details and Rationality

We found a subtle trade-off between surface details and rationality. Higher surface details mean closer to the predicted 2.5D sketches, which is often accompanied by a decline in the rationality of the reconstruction, i.e., less conformity with the SMPL model prediction. As shown in Figure 15, when we attempt to obtain better surface details using a pixel-aligned feature extractor with higher resolution, very irrational conditions appear in the character's head. However, the reconstruction results are extremely close to the reconstructed 2.5D sketches (especially the normal map), and the surface has more details. On the other hand, when not relaxing the SMPL constraints, we can see that all the details we want to add through the multilevel network have turned into noise around the SMPL model, but the overall reconstruction results are around the SMPL model, and the proportions of the head are reasonable. In our approach, we aim to add as much surface detail as possible without losing rationality, and further ensuring rationality in some extreme cases with the fusion repair algorithm. Even so, one of the problems that need to be addressed in the future is how to ensure rationality while optimizing surface details.

## 6. Discussion

We propose a multi-level single-view human 3D reconstruction framework that utilizes SMPL features and 2.5D sketches features as priors to constrain the plausibility of the reconstruction results. Coarse IF-Net are used to add sufficient detail to the reconstruction results, which are further optimized and denoised through fine IF-Net. Then, robust, reasonable, and detailed reconstructions are achieved by using our fusion and repair algorithm to combine coarse and fine occupancy fields. The effectiveness of our method is demonstrated through comparative experiments and ablation studies.

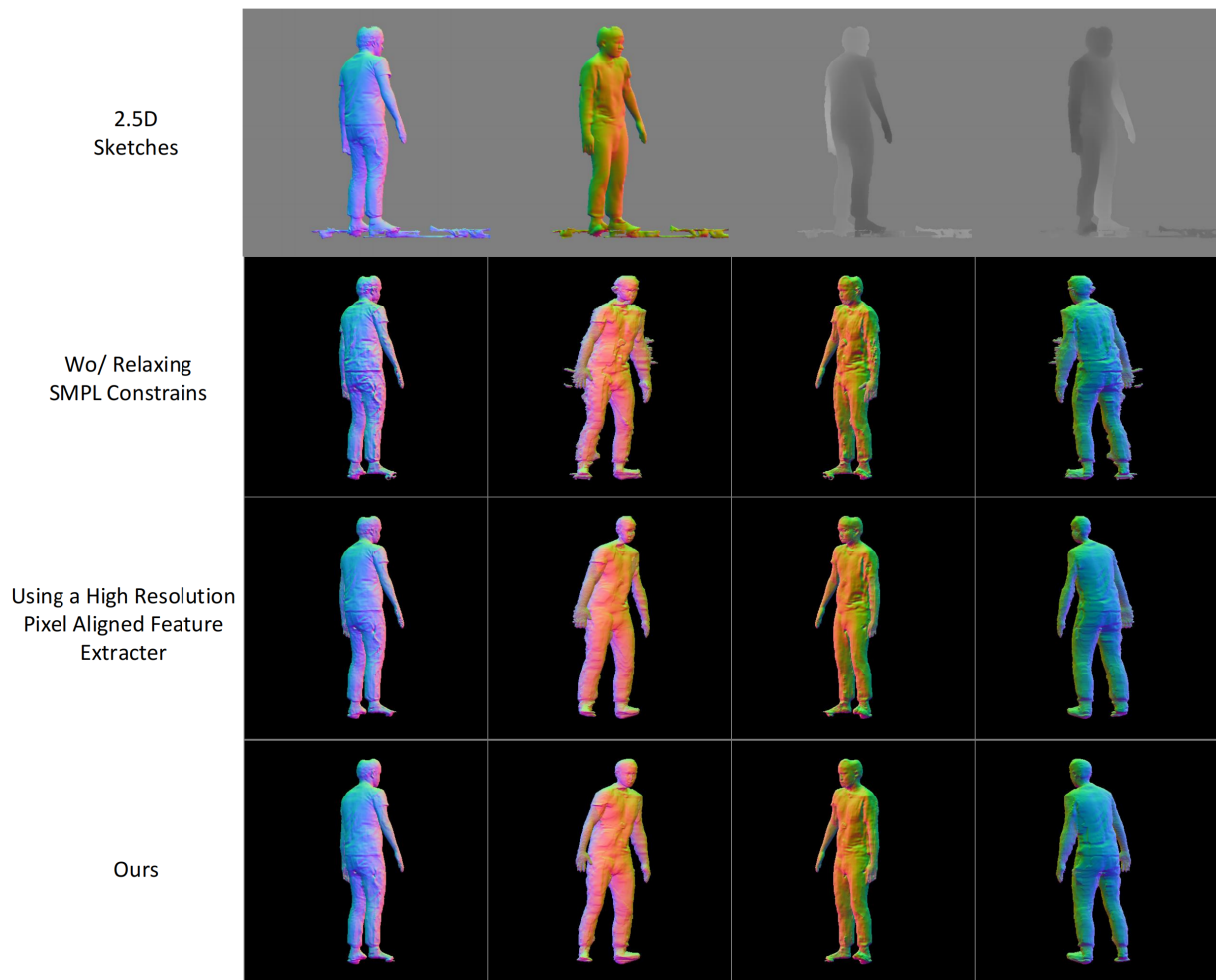
**Limitations and Future Work.** Our method is still based on a single image. If given a video sequence, one possible future research direction is to leverage temporal relationships to achieve faster and more consistent reconstructions.

Moreover, even we relax the SMPL constraints to incorporate more surface details, the quality of SMPL model predictions remains crucial to the reconstruction results. A flawed SMPL model prediction can lead to reconstruction failures, and local inaccuracies in SMPL predictions may result in deviations between the reconstructed output and the input images. Consequently, integrating a more robust and precise method for SMPL model prediction would significantly enhance the accuracy of our method.

Additionally, this paper employs a fusion and repair method to address the potential absence of limbs in fine reconstructions. Due to the rich details and diverse poses of limbs, especially hands, there are still issues with generating unreasonable results even in the state-of-the-art image generation field. Currently, digital human

technology and many explicit human body reconstruction techniques are emerging. High-quality implicit and explicit reconstructions can complement each other and serve as priors to enhance the reconstruction quality. For example, implicit function methods can be used to reconstruct clothing and wrinkles, while explicit models can be employed for reconstructing faces, hands, and feet, thus achieving more refined and reasonable reconstructions.





**Figure 15:** Trade-off between surface details and rationality

## References

- [APMTM19] ALLDIECK T., PONS-MOLL G., THEOBALT C., MAGNOR M.: Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2293–2303. [1](#)
- [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. [1, 3](#)
- [BTFB21] BECHTOLD J., TATARCHENKO M., FISCHER V., BROX T.: Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15880–15889. [3](#)
- [ÇAL\*16] ÇIÇEK Ö., ABDULKADIR A., LIENKAMP S. S., BROX T., RONNEBERGER O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19* (2016), Springer, pp. 424–432. [3](#)
- [CPMAMN22] CORONA E., PONS-MOLL G., ALENYÀ G., MORENO-NOGUER F.: Learned vertex descent: A new direction for 3d human model fitting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022). [1, 9](#)
- [CPW\*18] CHA Y.-W., PRICE T., WEI Z., LU X., REWKOWSKI N., CHABRA R., QIN Z., KIM H., SU Z., LIU Y., ET AL.: Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE transactions on visualization and computer graphics* 24, 11 (2018), 2993–3004. [1](#)
- [CXG\*16] CHOY C. B., XU D., GWAK J., CHEN K., SAVARESE S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14* (2016), Springer, pp. 628–644. [3](#)
- [FP09] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multi-view stereopsis. *IEEE transactions on pattern analysis and machine intelligence* 32, 8 (2009), 1362–1376. [1](#)
- [GFM\*19] GABEUR V., FRANCO J.-S., MARTIN X., SCHMID C., RO-

- GEZ G.: Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 2232–2241. 2, 3, 4
- [HCJS20] HE T., COLLOMOSSE J., JIN H., SOATTO S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems 33* (2020), 9276–9287. 2, 3
- [HTCH15] HUANG P., TEJERA M., COLLOMOSSE J., HILTON A.: Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. *ACM Transactions on Graphics (ToG)* 34, 2 (2015), 1–14. 1
- [HTM19] HÄNE C., TULSIANI S., MALIK J.: Hierarchical surface prediction. *IEEE transactions on pattern analysis and machine intelligence* 42, 6 (2019), 1348–1361. 3
- [HXL\*20] HUANG Z., XU Y., LASSNER C., LI H., TUNG T.: Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3093–3102. 3
- [HZJ\*21] HONG Y., ZHANG J., JIANG B., GUO Y., LIU L., BAO H.: Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 535–545. 3
- [JCSN20] JINKA S. S., CHACKO R., SHARMA A., NARAYANAN P.: Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *2020 International Conference on 3D Vision (3DV)* (2020), IEEE, pp. 879–888. 3
- [KBH06] KAZHDAN M., BOLITHO M., HOPPE H.: Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing* (2006), vol. 7, p. 0. 3
- [LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics* 21, 4 (1987), 163–169. 2, 3
- [LMR\*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16. 1, 3, 4
- [MNSL22] MOON G., NAM H., SHIRATORI T., LEE K. M.: 3d clothed human reconstruction in the wild. In *European Conference on Computer Vision (ECCV)* (2022). 1, 9
- [MON\*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4460–4470. 3
- [MST\*21] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHY R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65, 1 (2021), 99–106. 3
- [MYR\*20] MA Q., YANG J., RANJAN A., PUJADES S., PONS-MOLL G., TANG S., BLACK M. J.: Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6469–6478. 8
- [OBB20] OSMAN A. A., BOLKART T., BLACK M. J.: Star: Sparse trained articulated human body regressor. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16* (2020), Springer, pp. 598–613. 3
- [OERF\*16] ORTS-ESCOLANO S., RHEMANN C., FANELLO S., CHANG W., KOWDLE A., DEGYAREV Y., KIM D., DAVIDSON P. L., KHAMIS S., DOU M., ET AL.: Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology* (2016), pp. 741–754. 1
- [OHT\*20] ONIZUKA H., HAYIRCI Z., THOMAS D., SUGIMOTO A., UCHIYAMA H., TANIGUCHI R.-I.: Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6011–6020. 2
- [PFS\*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 165–174. 3
- [PMPHB17] PONS-MOLL G., PUJADES S., HU S., BLACK M. J.: Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–15. 1
- [RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18* (2015), Springer, pp. 234–241. 3
- [ROUG17] RIEGLER G., OSMAN ULUSOY A., GEIGER A.: Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 3577–3586. 3
- [SBR22] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision* (2022). 3
- [SHN\*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 2304–2314. 2, 3, 6, 8, 9
- [SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 84–93. 2, 3, 4
- [SYZR21] SU S.-Y., YU F., ZOLLHOEFER M., RHODIN H.: A-nerf: Surface-free human 3d pose refinement via neural rendering. *arXiv preprint arXiv:2102.06199* (2021). 3
- [TDB17] TATARCHENKO M., DOSOVITSKIY A., BROX T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2088–2096. 3
- [TZEM17] TULSIANI S., ZHOU T., EFROS A. A., MALIK J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2626–2634. 3
- [VCR\*18] VAROL G., CEYLAN D., RUSSELL B., YANG J., YUMER E., LAPTEV I., SCHMID C.: Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 20–36. 1, 3
- [WLZ\*18] WANG T.-C., LIU M.-Y., ZHU J.-Y., TAO A., KAUTZ J., CATANZARO B.: High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8798–8807. 5
- [WSK\*15] WU Z., SONG S., KHOSLA A., YU F., ZHANG L., TANG X., XIAO J.: 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1912–1920. 3
- [WZL\*18] WANG N., ZHANG Y., LI Z., FU Y., LIU W., JIANG Y.-G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 52–67. 3
- [XWC\*19] XU Q., WANG W., CEYLAN D., MECH R., NEUMANN U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems* 32 (2019). 3
- [XYTB22] XIU Y., YANG J., TZIONAS D., BLACK M. J.: Icon: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), IEEE, pp. 13286–13296. 2, 3, 4, 5, 9

- [YSR\*20] YAO Y., SCHERTLER N., ROSALES E., RHODIN H., SIGAL L., SHEFFER A.: Front2back: Single view 3d shape reconstruction via front to back prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 531–540. [3](#)
- [YZG\*21] YU T., ZHENG Z., GUO K., LIU P., DAI Q., LIU Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 5746–5756. [8](#)
- [ZTZ\*21] ZHANG H., TIAN Y., ZHOU X., OUYANG W., LIU Y., WANG L., SUN Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 11446–11456. [4](#)
- [ZYLD21] ZHENG Z., YU T., LIU Y., DAI Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence* *44*, 6 (2021), 3170–3184. [2](#), [3](#), [4](#), [9](#)
- [ZYW\*19] ZHENG Z., YU T., WEI Y., DAI Q., LIU Y.: Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7739–7749. [1](#), [3](#)