ROBUST-PIFU: ROBUST PIXEL-ALIGNED IMPLICIT FUNCTION FOR 3D HUMAN DIGITALIZATION FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing methods for 3D clothed human digitalization perform well when the input image is captured in ideal conditions that assume the lack of any occlusion. However, in reality, images may often have occlusion problems such as incomplete observation of the human subject's full body, self-occlusion by the human subject, and non-frontal body pose. When given such input images, these existing methods fail to perform adequately. Thus, we propose Robust-PIFu, a pixel-aligned implicit model that capitalized on large-scale, pretrained latent diffusion models to address the challenge of digitalizing human subjects from non-ideal images that suffer from occlusions.

Robust-PIfu offers four new contributions. Firstly, we propose a 'disentangling' latent diffusion model. This diffusion model, pretrained on billions of images, takes in any input image and removes external occlusions, such as inter-person occlusions, from that image. Secondly, Robust-PIFu addresses internal occlusions like self-occlusion by introducing a 'penetrating' latent diffusion model. This diffusion model outputs multi-layered normal maps that by-pass occlusions caused by the human subject's own limbs or other body parts (i.e. self-occlusion). Thirdly, in order to incorporate such multi-layered normal maps into a pixel-aligned implicit model, we introduce our Layered-Normals Pixel-aligned Implicit Model, which improves the structural accuracy of predicted clothed human meshes. Lastly, Robust-PIFu proposes an optional super-resolution mechanism for the multi-layered normal maps. This addresses scenarios where the input image is of low or inadequate resolution. Though not strictly related to occlusion, this is still an important subproblem. Our experiments show that Robust-PIFu outperforms current SOTA methods both qualitatively and quantitatively. Our code will be released to the public.

1 INTRODUCTION

3D Human Digitalization is the problem of reconstructing or generating a 3D model (e.g. a mesh) of a human subject from inputs such as a single image of that human subject. Recently, Wang et al. (2023b) showed that existing 3D Human Digitalization methods, such as ICON (Xiu et al., 2022) and PIFuHD (Saito et al., 2020), face severe difficulties when given input images that have incomplete observation (i.e. missing pixels) of the human subject's full body (like in Fig. 1c-d). These missing pixels often occur when there are multiple human subjects in the same input image (the subjects occlude one another). Wang et al. (2023b) suggested a method to address this incomplete observation problem, but they assumed that a groundtruth SMPL-X mesh (Pavlakos et al., 2019) is available and provided. A groundtruth SMPL-X mesh is a clothless and hairless 3D model that has both the actual 3D body pose and 3D body shape (excluding clothes and hair) of a human subject. Assuming that a groundtruth SMPL-X mesh is available greatly simplifies the problem of incomplete observation, but it is practically unrealistic. In most situations, the only input that is available is the input image itself.

Thus, we decided to address this problem. To this end, we introduce Robust-PIFu, a pixel-aligned implicit model that capitalized on large-scale, pretrained latent diffusion models to circumvent wide-

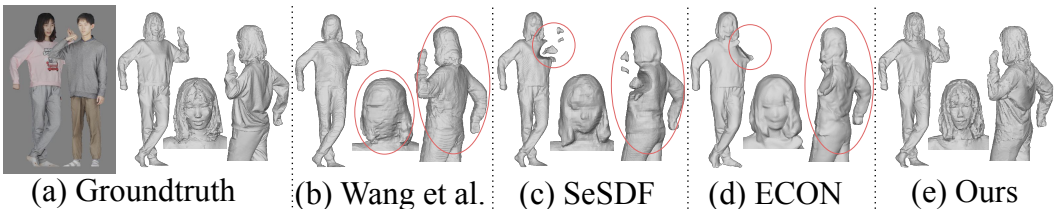


Figure 1: Results of SOTA and RobustPIFu on input images (in color) that have external occlusions.

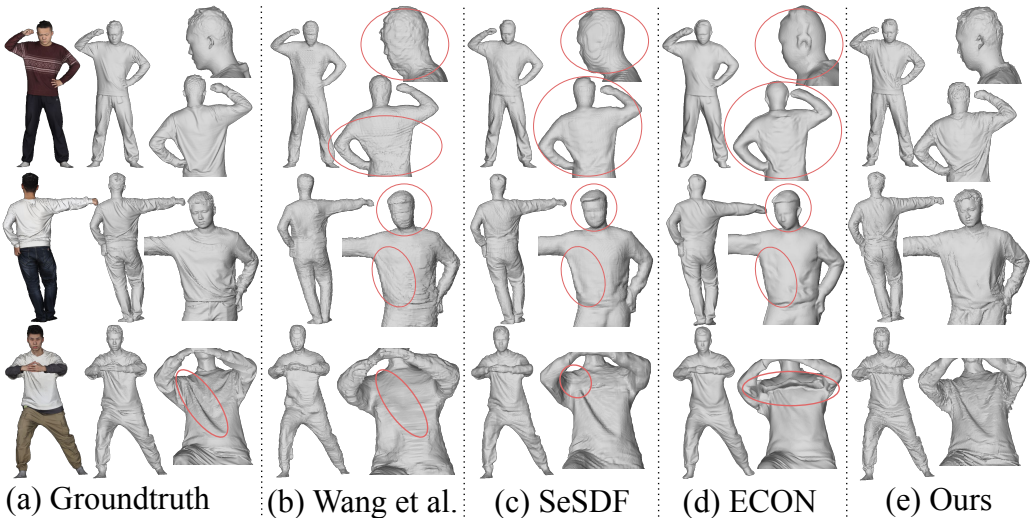


Figure 2: Results on input images (in color) that have no occlusion (1st row), internal occlusion due to non-frontal pose (2nd row), and internal occlusion due to self-occlusion (3rd row)

ranging occlusion problems (including incomplete observation) that may or may not occur in an input image.

Our work is inspired by the observations documented in a recent work, Zero-1-to-3 (Liu et al., 2023b). Zero-1-to-3 found that a large-scale, pretrained latent diffusion model, when given an image of an object as an input, can be controlled to synthesize a novel view (image) of that object from any specified viewpoint. This is significant as the latent diffusion model has demonstrated a 3D conceptual understanding of the object by freely generating new images of the same object from different viewpoints. To us, this is an intriguing finding as a novel view is, in all cases under this context, a view that contains unseen regions of an object. This bears similarities to our own problem (i.e. occlusion in input images), which pertains to uncovering unseen regions that have been occluded. Hence, we would like to find out if we can design a similar mechanism to control the pretrained latent diffusion model but instead of synthesizing novel views, we want the model to synthesize views that uncover and resolve any external and internal occlusions.

We classify occlusion in an input image into **external and internal occlusions**. We define external occlusions as occlusions that arise due external factors unrelated to the pictured human subject. Examples include an object occluding the human subject, or another human occluding the human subject that we are interested in (also called **inter-person occlusion**). On the other hand, we define internal occlusions as occlusions caused by factors that can be controlled by the human subject. Examples include self-occlusion (subject uses his/her limbs to cover a part of his/her body) or occlusion caused by the human subject not directly facing the camera (i.e. **non-frontal body pose**).

Firstly, to address external occlusions, our RobustPIFu introduces a pretrained latent diffusion model that is controlled to 'disentangle' human subject(s) from a given input image. This diffusion model isolates the person of interest from any extraneous objects or other humans, thereby resolving external occlusions. This diffusion model is the first contribution of Robust-PIFu.

Secondly, to tackle internal occlusions, we propose another pretrained latent diffusion model that is controlled to predict multi-layered normal maps at different camera angles. Regardless of whether the human subject tried to self-occlude or face away from the camera, these multi-layered normal

maps will show the full geometry of the human subject's body. This is the second contribution of Robust-PIFu.

Thirdly, in order to incorporate such multi-layered normal maps into a pixel-aligned implicit model, we introduce our Layered-Normals Pixel-aligned Implicit Model. Layered-Normals Pixel-aligned Implicit Model uses two different mechanisms (introduced later) to incorporate the multi-layered normal maps and serves to improve the structural accuracy of generated clothed human meshes.

Lastly, besides external and internal occlusions, another common problem in input images is the low or inadequate resolution of the pictured human subject. This can occurs when the human subject is far away from the camera or when the camera is set to capture more than one human subject. In such cases, existing methods will not be able to reconstruct a clothed human mesh with high-resolution details. In order to overcome this, we designed an optional mechanism to control a pretrained latent diffusion model for normal map super-resolution. This allows Robust-PIFu to create meshes with high-resolution details despite a low-resolution input image.

2 Related Works

2.1 Single-view Clothed Human Digitalization

Single-view clothed human digitalization refers to the task of retrieving or reconstructing a 3D model (e.g. mesh) of a clothed human body from a single RGB image. An important class of deep learning methods that have consistently attained state-of-the-art results for single-view clothed human digitalization is the pixel-aligned implicit models. These models, which include ECON (Xiu et al., 2023), ICON (Xiu et al., 2022), IntegratedPIFu (Chan et al., 2022), and PIFuHD (Saito et al., 2020), work by learning an implicit function that represents the enclosed surface of a clothed human body. During testing, the learned implicit function is sampled using a 3D grid of evenly-spaced sample points. For each sample point, the implicit function returns/predicts an occupancy label, which indicates if the sample point is considered 'inside' or 'outside' of a clothed human body's surface. Once a 3D grid of predicted occupancy labels is computed, a mesh that resembles a clothed human body can be extracted from this grid by using the Marching Cubes algorithm (Lorensen & Cline, 1987). For more information on pixel-aligned implicit models, refer to Annex A.18.1.

2.2 Robustness in Pixel-aligned Implicit Models

Pixel-aligned implicit models' robustness to occlusions is an important topic that has garnered attention from researchers. For example, ECON (Xiu et al., 2023), which is a pixel-aligned implicit model, has proposed mechanisms to improve robustness with respect to inter-person occlusions. One of the mechanisms proposed by ECON is to train a shape completion module with a parametric human body model (i.e. a SMPL-X mesh) and a set of randomly masked depth maps. In addition, the work by Wang et al. (2023b) proposes a new pixel-aligned implicit model that also aims to be robust to inter-person occlusions in a given input image. The proposed pixel-aligned implicit model infers the occluded or missing information of a human subject's body by incorporating a SMPL-X mesh, a 2D GAN, and a 3D GAN into its pipeline.

However, there are a number of problems with how these existing models address robustness questions. Firstly, both ECON and the work by Wang et al. assume that a groundtruth or a very accurate SMPL-X mesh is available as an input. In situations where there are occlusions or missing pixels in a given input image, it is unlikely that a very accurate SMPL-X mesh can be predicted from that input image. Thus, these models may not be practical in real-life use cases.

Secondly, even if we assume groundtruth SMPL-X meshes are given, the clothed human meshes reconstructed by ECON and the work by Wang et al. often suffer from poor structural accuracy and imprecise details when there are inter-person occlusions in the given input image (See Fig. 1b,d).

Lastly, other than inter-person occlusions (which is a type of external occlusion), there are other types of occlusions like internal occlusions. Internal occlusions, such as self-occlusion, are not well-addressed by existing methods yet.

Hence, we propose Robust-PIFu in order to address both external and internal occlusions effectively.

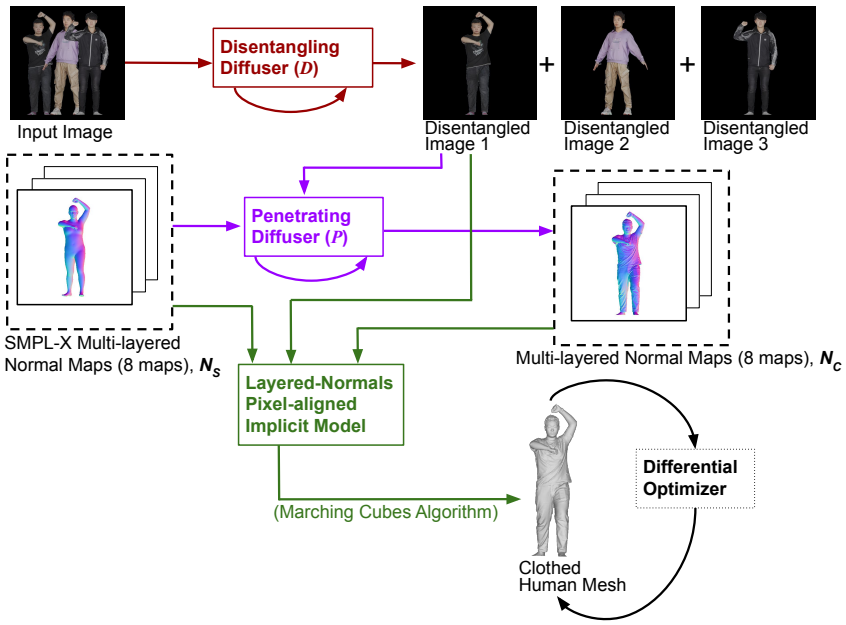


Figure 3: Overview of our Robust-PIFu. D first 'disentangled' the human subjects into separate images. Then, P predicts multi-layered normal maps from each disentangled image. The maps are used by our Layered-Normals Pixel-aligned Implicit Model to construct a clothed human mesh.

2.3 Large-scale Latent Diffusion Models

Recently, large-scale 2D diffusion models, such as DALL-E (Ramesh et al., 2021), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022), have showcased their capacity to grasp an extensive range of visual concepts from Internet-scale image datasets. Consequently, an increasing number of researchers are harnessing these models' rich priors about our 3D world and utilizing them in 3D tasks (e.g. (Poole et al., 2022; Lin et al., 2023; Liu et al., 2023b;a; Chen et al., 2024; Lee et al., 2024; Ho et al., 2024)).

One notable example is Zero-1-to-3 (Liu et al., 2023b). Zero-1-to-3 illustrated that a latent diffusion model, pretrained on extensive internet image data, can be manipulated to produce a fresh perspective of an object from any specified viewpoint. To achieve this, the pretrained model first undergoes fine-tuning with an input RGB view of the object, a relative camera transformation, and a transformed RGB view of the same object. Once fine-tuned, the model, when provided with an input view of the object, can generate numerous novel views of that object. These new views are then utilized to reconstruct a 3D mesh of the object using Score Jacobian Chaining (SJC) (Wang et al., 2023a). Zero-1-to-3 demonstrated superior performance compared to state-of-the-art (SOTA) results in 3D object reconstruction. This finding not only shows that vast internet image data can be leveraged on to solve a 3D reconstruction task, but it also demonstrates that a pre-trained latent diffusion model is able to gain a 3D conceptual understanding of an object just by processing a 2D image of that object. (For more information on Zero-1-to-3, refer to Annex A.18.2.)

We wish to harness and capitalize on the 3D conceptual understanding attained by a pre-trained latent diffusion model as well. We hypothesize that a 3D conceptual understanding of a human will be useful in identifying and removing occlusions that are present in a 2D image of that human. Just as the Zero-1-to-3 uses a pre-trained latent diffusion model to uncover an unseen view (i.e. a novel view) of an object, we wish to use a pre-trained latent diffusion model to uncover unseen regions (i.e. occluded regions) of a human subject. In other words, we aim to use the latent diffusion model to resolve any external and internal occlusions that may be present in a given input image.

3 Method

We show an overview of Robust-PIFu in Fig. 3. First, Robust-PIFu is given an input image that may or may not have any occlusion. This given input image is passed to Disentangling Diffuser

(D), which is a latent diffusion model that will remove any external occlusion (e.g. inter-person occlusions) from the input image. If there are inter-person occlusions, this means there are multiple subjects in the input image, and D will 'disentangle' the human subjects into separate images (i.e. disentangled images). Each disentangled image, now free of external occlusions, can then be easily used to predict a SMPL-X mesh (if the groundtruth SMPL-X mesh is not given).

Next, each of the disentangled images are separately fed into our second latent diffusion model (i.e. Penetrating Diffuser or P). For each disentangled image, P will generate 8 normal maps (also called **multi-layered normal maps** or N_C), which correspond to the different layers of the human subject at different camera angles. These 8 normal maps allow us to bypass both self-occlusion and non-frontal pose of the human subject in the given input image. In other words, P resolves internal occlusions.

Finally, N_C and the disentangled image are passed to our Layered-Normals Pixel-aligned Implicit Model, which is a pixel-aligned implicit model fitted with two new mechanisms. These mechanisms, which will be introduced later, enable multi-layered normal maps to be incorporated into a pixel-aligned implicit model. Layered-Normals Pixel-aligned Implicit Model produces a clothed human mesh, which is then refined by a differential optimizer using N_C.

In order to incorporate a SMPL-X mesh into our Robust-PIFu, we first extract multi-layered normal maps from the SMPL-X mesh (N_S). N_S is similar to N_C except that N_S pertains to a SMPL-X body, and N_C pertains to a clothed human body. N_S is then fed into both P and the Layered-Normals Pixel-aligned Implicit Model.

Also, Robust-PIFu also offers an optional diffuser - our Refining Diffuser (R). Our R will super-resolve N_C, allowing higher-resolution N_C to be used in the differential optimizer. However, the tradeoff of using R is increased inference time.

Now, we will elaborate on the four contributions of Robust-PIFu: 1. Disentangling Diffuser (Sect. 3.1) 2. Penetrating Diffuser (Sect. 3.2) 3. Layered-Normals Pixel-aligned Implicit Model (Sect. 3.3) 4. Refining Diffuser (Sect. 3.4).

3.1 Disentangling Diffuser (D)

As explained in Sect. 2.2, existing methods, like ECON and the work by Wang et al. (2023b), do not address external occlusion like inter-person occlusion well because these models assume that a very accurate SMPL-X mesh is available despite the occlusion. Moreover, even when this assumption is true, these models often fail to produce clothed meshes that have accurate structure and precise details.

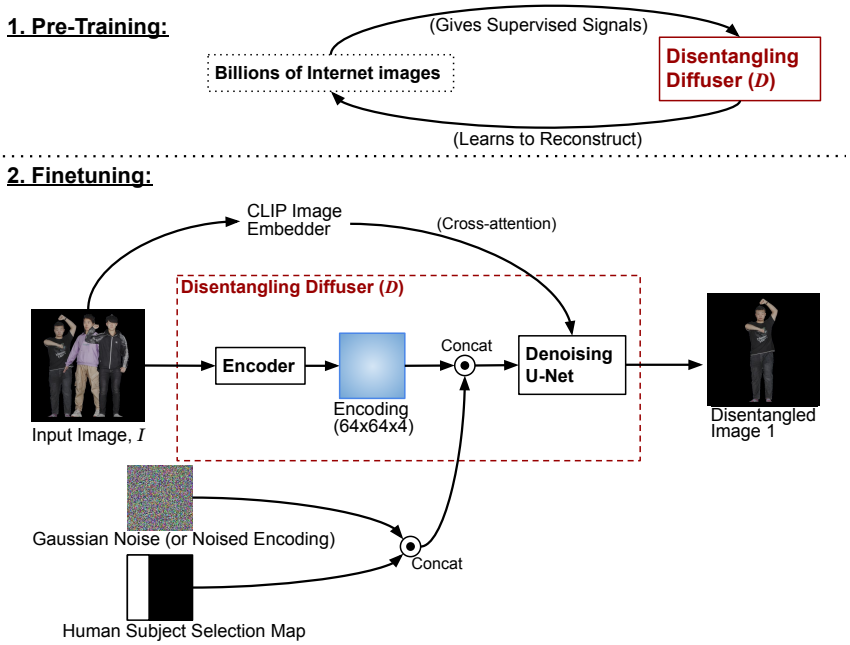


Figure 4: The Pre-training and Finetuning stages of Disentangling Diffuser (D).

To address external occlusion, we propose Disentangling Diffuser (D). Our D is shown in Fig. 4, and it capitalizes on the extensive visual concepts and knowledge learned by a large-scale latent diffusion model that has been pre-trained on billions of images. Similar to Zero-1-to-3 (Liu et al., 2023b), our D injects the pre-trained latent diffusion model with architectural modifications before a round of finetuning. Specifically, D aims to denoise a noised encoding that represents an image of a human subject. As shown in Fig. 4, the given input image (I) to D is an image of at most three human subjects. I is used as a conditional prior to D via two streams. First, we obtain the CLIP image embedding of I before using it for cross-attention in the denoising U-Net of the latent diffusion model. Second, I is encoded into an encoding and then channel-concatenated with a noised encoding before it is fed to the denoising U-Net. Also, in order to control which of the three human subjects to extract out, we introduce a Human Subject Selection Map, which is a binary map that has one out of three possible patterns (see Annex A.2 for more info). The denoising U-Net will output a denoised encoding that is then decoded into a 'disentangled' image. (Aside: D can deal with input images that contain more than three human subjects as well, see Annex A.2). A formal formulation of D is given in Eqn. 1.

$$I_i^d = D(I, H_i), \quad i \in \mathbb{Z}, \quad 0 < i \leq n \tag{1}$$

where I_i^d is the disentangled image of the i^{th} human subject, I is the input image, H_i is the i^{th} Human Subject Selection Map, and n is the maximum number of human subjects allowed in the input image to D. In our work, we set n to be 3.

To address a wide range of external occlusions, D is finetuned with I that is afflicted with different types of external occlusions, such as inter-person occlusion, object-caused occlusion, and poor lighting conditions (see Figs. 11 and 12). After finetuning, our D will be able to remove these various forms of external occlusions from any given input image I, as also shown in Figs. 11 and 12.

The training or finetuning process of D (as well as P and R) follows a typical latent diffusion model whereby the model learns to predict the noise added to the latent encodings at each time step, and the MSE between the predicted and actual noises is minimized. More training details in Annex A.14.

3.2 Penetrating Diffuser (P)

The disentangled image produced by D may still be affected by internal occlusions, such as self-occlusion and occlusion caused by non-frontal pose of the human subject. For example, a human subject may put his arm across this torso, obscuring the topology of his jacket. Similarly, a human subject that is not facing the camera directly may conceal the topology of his facial features. Internal occlusions like these will prevent a pixel-aligned implicit model from reconstructing a clothed human mesh that has an accurate and detailed topology. In order to bypass internal occlusions, we

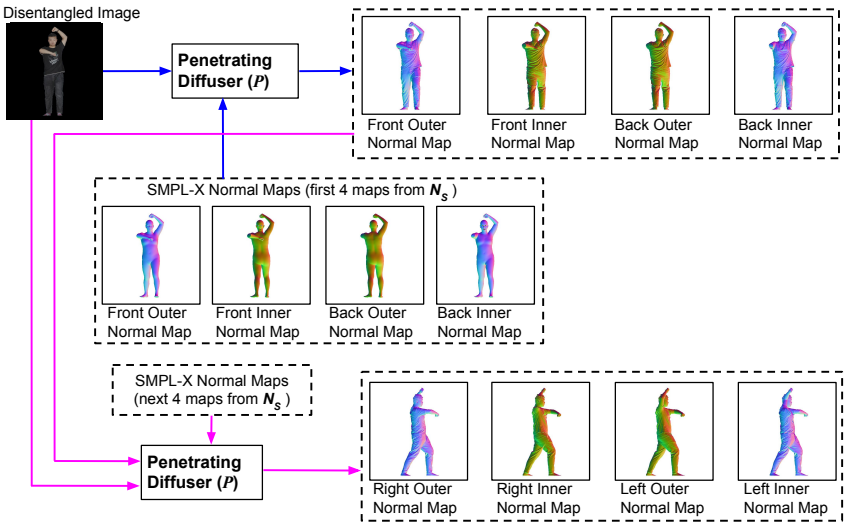


Figure 5: Penetrating Diffuser (P). In the 1st run (blue arrows), 4 normal maps from the front and back camera directions are produced. In the 2nd run (magenta arrows), a further 4 normal maps from the right and left camera directions are produced. These 8 normal maps constitute our N_C.

propose our Penetrating Diffuser (P), which predicts **multi-layered normal maps** of a clothed human body. These maps are also known as N_C, and they describe the topology of a human subject at different layers and from distinct camera directions (illustrated and explained in Annex A.4).

To design P, we again tap on the visual knowledge acquired by a large-scale pre-trained latent diffusion model. From the pre-trained model, we extend it by adding two new conditional priors (See Fig. 5). The first conditional prior is the disentangled image produced by D. This disentangled image is incorporated into P the same way I is incorporated into D (i.e. via cross-attention and channel concatenation). The second conditional prior is N_S, which refers to multi-layered normal maps that are extracted from a SMPL-X mesh. N_S is incorporated into P via channel concatenation only. Unlike D, P does not output a RGB image in each run. Instead, P outputs a **set of 4 normal maps** in each run. In total, as shown in Fig. 5, P will be run twice. In the first run, 4 normal maps from the front and back camera directions are produced. In the second run, these 4 normal maps are used as conditional priors to P, and a further 4 normal maps from the right and left camera directions will then be produced. These 8 normal maps form our N_C i.e. the multi-layered normal maps of the clothed human body. These maps reveal the topology of the human subject from different sides and different layers. This resolves internal occlusions like self-occlusion or non-frontal body pose. A formal formulation of P is given in Eqn. 2.

$$N_C^t = P(I_i^d, N_S^t, N_C^{t-1}), \quad t \in \{1, 2\} \tag{2}$$

where I_i^d is the disentangled image of the i^{th} human subject. When t = 1, N_C^t and N_S^t represent the first 4 maps from N_C and N_S respectively. When t = 2, N_C^t and N_S^t represent the next 4 maps from N_C and N_S respectively. N_C^0 is a set of zero-filled maps.

3.3 Layered-Normals Pixel-aligned Implicit Model

At this stage, we have obtained multi-layered normal maps (both N_S and N_C) that contain information of the human subject from different sides and layers. But a pixel-aligned implicit model is not designed to take in these layered normal inputs. Thus, we propose a new form of pixel-aligned implicit model called **Layered-Normals Pixel-aligned Implicit Model**, which is shown in Fig. 6

Layered-Normals Pixel-aligned Implicit Model incorporates multi-layered normal maps via two mechanisms or streams. In the first stream, we concatenate N_S, N_C, and the disentangled image together. The concatenated product is then fed into the encoder (i.e. stacked hourglass network) of our pixel-aligned implicit model.

In the second stream, we use N_C to form a **Multi-dimensional Layered Normals Grid**. This 3D grid is formed by aligning the eight 2D normal maps of N_C inside a 3D space, as shown in Fig. 15. Each element/position on the grid is a 24-length vector, which consists of a normal vector (3-length vector) taken from each of the 8 normal maps. The grid contains not only detailed information on the topology of the human subject, but also the structural information of the human subject in the form of a visual hull (shown in bottom right of Fig. 15). This visual hull narrows down the possible shape of the human subject's body as the surface of the human subject cannot lie outside of the visual hull. In other words, the visual hull acts like an initial, template 3D shape for the pixel-aligned implicit model to work on before predicting a final clothed human shape/mesh. To

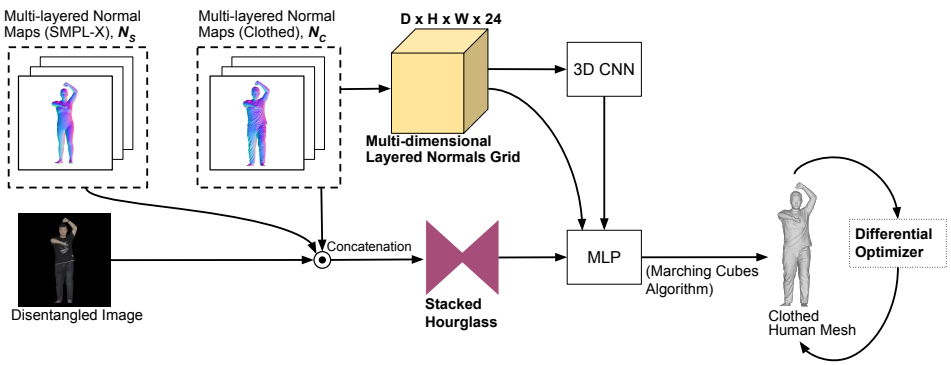


Figure 6: The final part of Robust-PIFu: Layered-Normals Pixel-aligned Implicit Model.

incorporate the Multi-dimensional Layered Normals Grid into the pixel-aligned implicit model, we process this multi-dimensional grid using a 3D-CNN (see Fig. 6). The features produced by the 3D-CNN are then fed into the MLP. In addition, we also allow for a skip connection that directly feeds the values from the grid into the MLP.

Via these two mechanisms, we allow information from both our N_S and N_C to be incorporated into a pixel-aligned implicit model in a principled and sensible way. Like a typical pixel-aligned implicit model, our Layered-Normals Pixel-aligned Implicit Model is trained by minimizing the MSE between the predicted and actual occupancy labels of the sample points (Saito et al., 2019). After training, our pixel-aligned implicit model model will produce a clothed human mesh using the Marching Cubes's algorithm. In addition, we provide a formal formulation of our Layered-Normals Pixel-aligned Implicit Model in the following equation:

$$F(I_i^d, N_S, N_C) = s(x, y, z), \quad s(x, y, z) \in [0, 1] \tag{3}$$

where I_i^d is the disentangled image of the i^{th} human subject, and $s(x, y, z)$ denotes the predicted implicit function that represents a 3D clothed human body.

The clothed human mesh is then refined by a differential optimizer, as is done in ICON (Xiu et al., 2022). The differential optimizer works by computing an error (e.g. L1 loss) between a predicted normal map and the surface normals of the clothed human mesh. The mesh is then deformed or adjusted such that the error is minimized. More than one predicted normal map can be used in this refining process. For Robust-PIFu, our N_C (i.e. 8 maps) is used as the predicted normal maps.

3.4 Refining Diffuser (R)

Our differential optimizer uses N_C to refine the meshes. In order to improve the refinement process, we can super-resolve N_C using our optional Refining Diffuser (R). The differential optimizer can then use the super-resolved N_C, rather than the original N_C, during the refinement process. R is a latent diffusion model that takes in N_C as a conditional prior and outputs a super-resolved N_C, as shown in Fig. 16. R addresses scenarios where the given input image is of low resolution (resulting in N_C having low resolution as well) or when an application requires clothed human meshes that are of a higher than usual resolution. However, the use of R will also result in a noticeably increased inference time. There is thus a 'inference time vs resolution' trade-off. More details on our R can be found in Annex A.7. In reality, inference time may be important to different applications. Thus, to ensure fairness, we did not use R to produce our results, but we demonstrate the usefulness of R in Fig. 17.

4 Experiments

4.1 Datasets

To train pixel-aligned implicit models (both ours and the existing models), 3D scans/data of human subjects are required. For our experiments, we use the THuman2.0 dataset (Yu et al., 2021) to train both our models and the existing models. The THuman2.0 dataset contains 526 3D full-body scans (i.e. meshes) of real-life human subjects. We use a 80-20 train-test split. For each 3D scan, we render RGB images at 10 degree intervals (azimuth rotation) using a weak-perspective camera.

Prior to training, our D, P, and R are already pre-trained with the same Internet-scale image datasets (Schuhmann et al., 2022) (>2 billion images) used by Stable Diffusion (Rombach et al., 2022).

In addition, we use BUFF dataset (Zhang et al., 2017) and MultiHuman dataset (Zheng et al., 2021) to evaluate the models. None of the models is trained using these two datasets. For BUFF dataset, we followed IntegratedPIFu (Chan et al., 2022) and performed systematic sampling (based on sequence number) on the dataset. This gives us a total of 101 human meshes that we use for evaluation. The purpose of using systematic sampling is to avoid meshes that have both the same human subject and the same pose. For MultiHuman dataset, all single human scans are used (see more in Annex A.11).

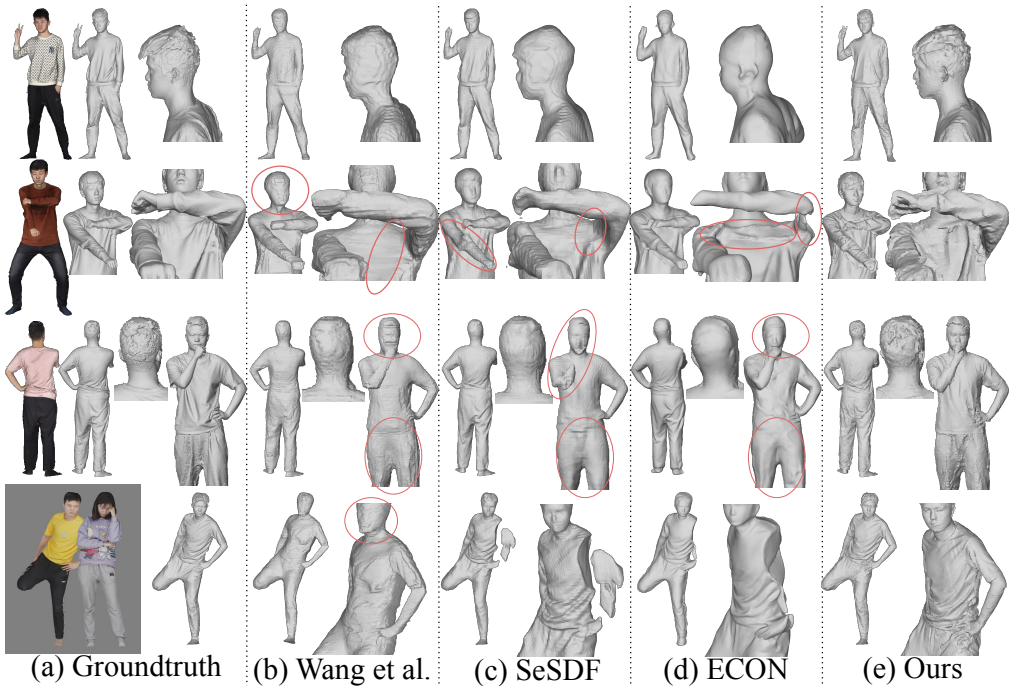


Figure 7: Results on input images (in color) that have no occlusion (1st row), internal occlusions due to self-occlusion (2nd row) and non-frontal pose (3rd row), and external occlusion (4th row)

Table 1: Quantitative Evaluation between SOTA and our Robust-PIFu.

	THuman2.0		BUFF		MultiHuman	
Methods	CD (10^{-5})	P2S (10^{-5})	CD (10^{2})	P2S (10^{2})	CD (10^{-5})	P2S (10^{-5})
ICON	9.947	9.014	8.659	7.892	15.76	13.39
ECON	11.17	9.090	9.064	9.750	14.69	13.13
D-IF	10.24	9.137	9.136	7.644	16.71	13.78
SeSDF	8.941	8.441	6.582	6.947	14.00	12.30
Wang et al.	8.944	8.596	9.072	9.881	21.01	21.64
Robust-PIFu (Ours)	**8.426**	**8.000**	**6.127**	**6.525**	**13.38**	**11.66**

4.2 Comparison with State-of-the-art

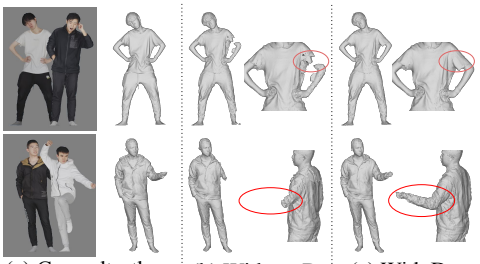
We compared Robust-PIFu against other existing models on single-view clothed human reconstruction. The models we compared with include ICON (Xiu et al., 2022), ECON (Xiu et al., 2023), D-IF (Yang et al., 2023), SeSDF (Single-view version) (Cao et al., 2023), and the work by Wang et al. (2023b). These models are selected not only because they achieved SOTA performances but also because they are designed to be robust to occlusions (e.g. inter-person occlusion, missing pixels in input image) and other various challenging conditions like complex human poses. In our quantitative evaluation, we followed (Cao et al., 2023) and used Chamfer Distance (CD) and Point-to-Surface (P2S) as metrics. CD is the bidirectional point-to-surface distances between GT and reconstructed meshes. P2S is the unidirectional variant of CD.

Qualitative Evaluation We evaluate the models qualitatively in Fig. 1, Fig. 2, and Fig. 7. Fig. 1 shows how the models react to external occlusions, while Fig. 2 pertains to internal occlusions. Fig. 7 shows more examples both with and without occlusion. From the figures, we observe that the existing models, unlike our Robust-PIFu, struggle to address external and internal occlusions and fail to produce meshes that have both an accurate structure and precise details. In Fig. 10, we show our results on real Internet images from Adobe Stock. More results in Annex A.8, A.9, and A.10.

Quantitative Evaluation We show our quantitative results in Tab. 1, and it shows that Robust-PIFu is able to outperform the existing models in all metrics for all three datasets.

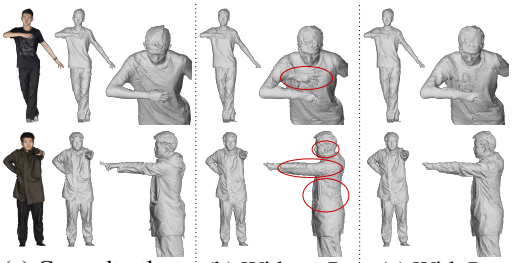
4.3 Ablation Studies

Evaluation of our Disentangling Diffuser (D) and its Effectiveness against External Occlusions Our Disentangling Diffuser (D) is designed to deal with external occlusions. In order to evaluate its



(a) Groundtruth (b) Without D (c) With D

Figure 8: Ablation of using and not using D.



(a) Groundtruth (b) Without P (c) With P

Figure 9: Ablation of using and not using P.

Table 2: Ablation study of Layered-Normals Pixel-aligned Implicit Model.

Methods	THuman2.0		BUFF		MultiHuman	
	CD (10^{-5})	P2S (10^{-5})	CD (10^{2})	P2S (10^{2})	CD (10^{-5})	P2S (10^{-5})
Robust-PIFu w/o Concat w/o Multi-D Grid	8.928	8.433	6.574	6.939	13.98	12.26
Robust-PIFu w/o Multi-D Grid	8.551	8.132	6.183	6.656	13.60	11.85
Robust-PIFu	**8.426**	**8.000**	**6.127**	**6.525**	**13.38**	**11.66**

effectiveness, we use given input images that have external occlusions and compare the 3D clothed human meshes that are produced by our RobustPIFu when D is used and when D is not used in Fig. 8. The results clearly show how D enables our RobustPIFu to produce unbroken, human-like meshes.

Evaluation of our Penetrating Diffuser (P) and its Effectiveness against Internal Occlusions Our Penetrating Diffuser (P) is designed to deal with internal occlusions. To evaluate its effectiveness, we use given input images that have internal occlusions and compare the 3D clothed human meshes that are produced by our RobustPIFu when P is used and when P is not used in Fig. 9. The results show that P is pivotal for constructing regions that are occluded or unseen in input images. See Annex A.4 for more ablation studies pertaining to P.

Evaluation of our Layered-Normals Pixel-aligned Implicit Model and its two Mechanisms In order to incorporate layered normal maps, we proposed using two mechanisms that modify the architecture of a pixel-aligned implicit model: 1. Concatenation of N_S, N_C, and the disentangled image 2. Multi-dimensional Layered Normal Grid. Here, we evaluate the impact of these two mechanisms in Tab. 2. The table shows both mechanisms considerably improved the structural accuracy of the clothed meshes.

5 Limitation and Conclusion

Limitation In Annex A.12, we address concerns pertaining to the computation time required by our Robust-PIFu. We do so by comparing Robust-PIFu with ECON (Xiu et al., 2023) in terms of computation time. See Annex A.13 for other limitations.

Conclusion We have proposed Robust-PIFu, a pixel-aligned implicit model that capitalized on large-scale pretrained latent diffusion models to tackle external and internal occlusions in given input images. Firstly, Robust-PIFu addresses external occlusion by proposing a Disentangling Diffuser that eliminates a range of external occlusions from an input image. Secondly, we introduce a Penetrating Diffuser that generates multi-layered normal maps to tackle internal occlusions. Thirdly, Robust-PIFu proposes a Layered-Normals Pixel-aligned implicit model that incorporates multi-layered normal maps into a pixel-aligned implicit model via two distinct streams. Finally, Robust-PIFu introduced an optional Refining Diffuser that performs super-resolution of normal maps, thereby allowing a high-resolution clothed human mesh to be produced even when given a low-resolution input image.

REFERENCES

Yukang Cao, Kai Han, and Kwan-Yee K Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4647–4657, 2023.

Kennard Yanting Chan, Guosheng Lin, Haiyu Zhao, and Weisi Lin. Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pp. 328–344. Springer, 2022.

Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail. *arXiv preprint arXiv:2403.12028*, 2024.

I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 538–549, 2024.

Jungeun Lee, Sanghun Kim, Hansol Lee, Tserendorj Adiya, and Hwasup Lim. Pidiffu: Pixel-aligned diffusion model for high-fidelity clothed human reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5172–5181, 2024.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems*, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023b.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pp. 483–499. Springer, 2016.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2304–2314, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 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*, pp. 84–93, 2020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023a.

Junying Wang, Jae Shin Yoon, Tuanfeng Y Wang, Krishna Kumar Singh, and Ulrich Neumann. Complete 3d human reconstruction from a single incomplete image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8748–8758, 2023b.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13286–13296. IEEE, 2022.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 512–523, 2023.

Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9122–9132, 2023.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021.

Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6239–6249, 2021.

A APPENDIX

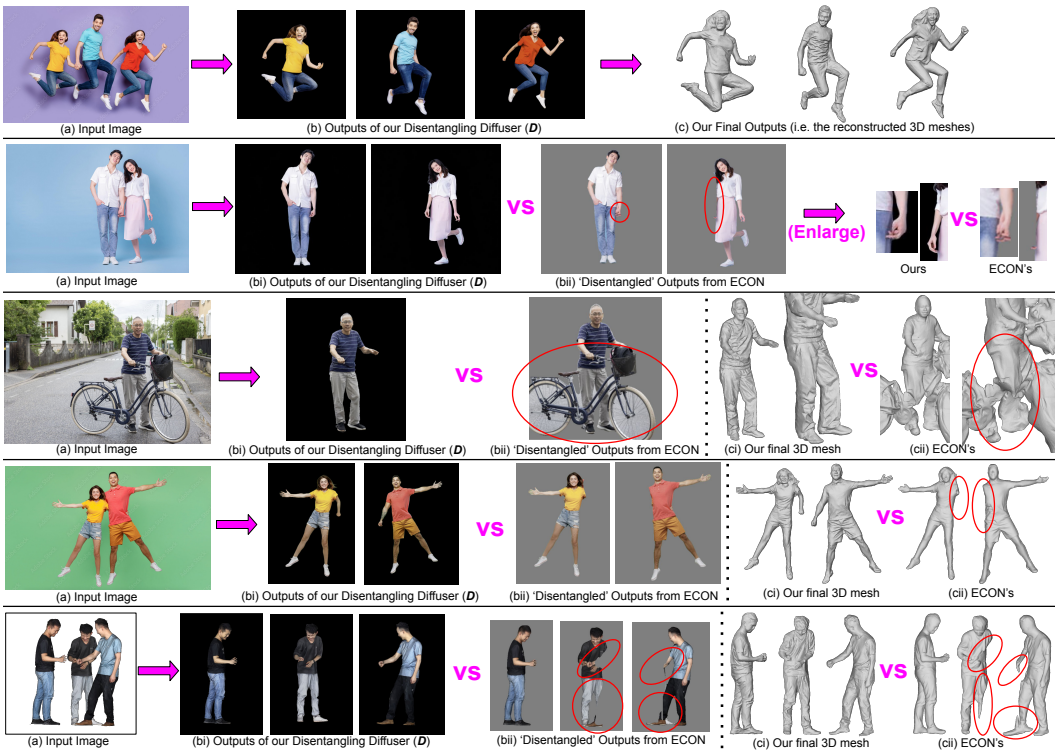


Figure 10: (Prior to the Rebuttal, this figure was placed in the Main Paper instead of the Appendix.) The first four rows pertain to our results on real images sourced from Adobe Stock. The last row pertains to our results on an image that is rendered from a group of humans that are physically in the same location during the scanning process.

A.1 PERFORMANCE OF DISENTANGLING DIFFUSER (D) FOR DIFFERENT TYPES OF EXTERNAL OCCLUSION

In Fig. 11 and Fig. 12, we show the outputs of D under different types of external occlusions.

A.2 HOW DISENTANGLING DIFFUSER (D) DEALS WITH INPUT IMAGES WITH LESS OR MORE THAN THREE HUMAN SUBJECTS

The Human Subject Selection Map is a 64x64x1 binary map that consists of zeros and ones. It has three possible patterns. In pattern 1, it is filled with all zeros except for the first (leftmost) 21 columns of the map. In pattern 2, it is filled with zeros except for the columns between the 21st and the 42nd row. In pattern 3, it is filled with zeros except for the last 22 columns. Pattern 1 indicates to D that the leftmost human subject in the given input image is to be extracted. Pattern 2 indicates the middle human subject, and Pattern 3 indicates the rightmost human subject. It is important to point out that the leftmost human subject does not need to occupy the leftmost 1/3 of the given input image. The subject can be positioned anywhere on the image, so long as it is the **leftmost** person in the given input image. This applies to the middle and rightmost human subjects as well.

If we feed in Pattern 1, Pattern 2, and Pattern 3 into D separately, then we will obtain 3 different output images. However, there may be situations where there are less than three subjects in the given input image. As shown in Fig. 12, when there are less than three subjects in the given input image, then one or more of the output images from D will be blank. Specifically, if there are only two subjects, then the output image generated using Pattern 3 will be blank. If there are only 1 subject, then both output images generated using Pattern 2 and Pattern 3 will be blank.

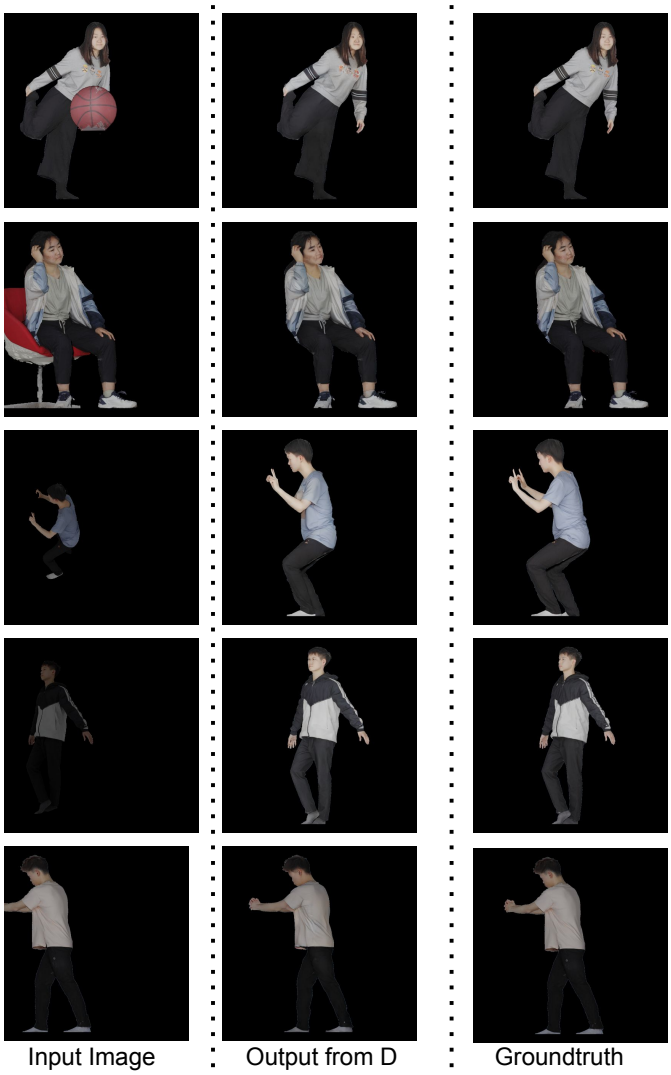


Figure 11: Results of D in different types of External Occlusion

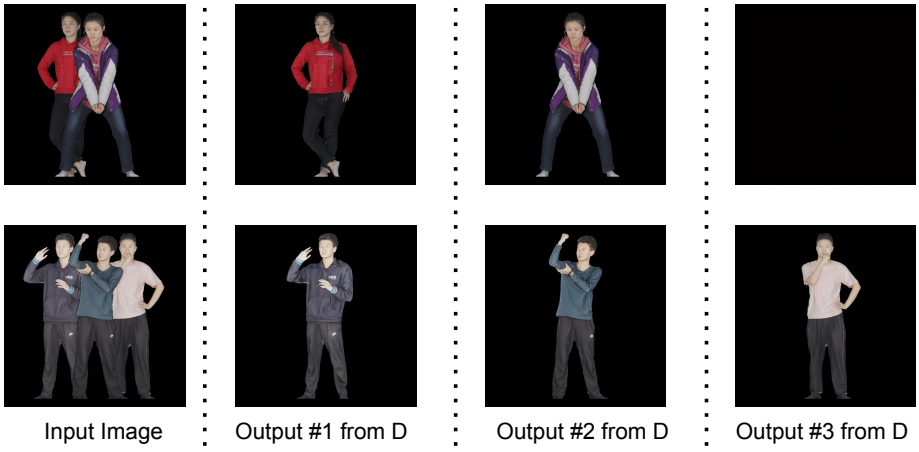


Figure 12: Results of D in Inter-person Occlusions.

On the other hand, if there are more than three human subjects (e.g. 5) in the given input image, we use an off-the-shelf human instance segmentation model to detect the rough positions of the 5 subjects and then split the image into two smaller images that contain 3 and 2 subjects respectively.

The two smaller images are then separately passed into our proposed model pipeline (starting with D). An actual example is given in the first two rows of Fig. 20.

A.3 How D deals with Vertically Occluded Humans

Our paper, like other SOTA papers like ECON Xiu et al. (2023) and the work by Wang et al. (2023b), show primarily "horizontally occluded human(s)" instead of "vertically occluded human(s)". Vertically occluded humans refer to images where the top or bottom (rather than left or right) of a human subject is occluded by another human subject or an object.

It is not true that our D cannot deal with vertically occluded human(s). During training, we can simply give pattern 1 (Human Subject Selection Map) to our D and ask it to produce the topmost human subject in the image of vertically occluded human(s) during training, and it would work well during testing because latent diffusion models do not fail just because the objects of interest in an image is arranged from top-to-bottom instead of from left-to-right.

The paragraph above assumes that an image of "vertically occluded human(s)" features humans that are perfectly aligned in a vertical line. Because if the humans are not aligned exactly as such, then our proposed method will already be able to address that image without additional training. In reality, we find this possibility of perfect alignment to be very rare. Most papers do not consider images of such perfectly aligned "vertically occluded human(s)". For example, both ECON and the work by Wang et al. do not feature humans that are perfectly aligned in a vertical line. In fact, both of them use only images with "horizontally occluded human(s)", just as we did in our paper.

A.4 The 8 Normal Maps (Multi-layered Normal Maps) generated by Penetrating Diffuser (P) and More Results from P

The multi-layered normal maps consist of 8 normal maps. We illustrate how these 8 normal maps are rendered in Fig. 13. The groundtruth 3D clothed human mesh is rendered from 4 different camera directions (front, back, right, and left). For each camera direction, we are interested not only in the first normal vector encountered by a camera ray (horizontal arrow in the figure) but also the second normal vector as well. The first normal vectors are placed in the 'Outer' normal map while the second normal vectors are placed in the 'Inner' normal map. In total, we will obtain 8 normal maps, and these maps reveal the topology of the human subject at different layers (i.e. inner and outer) and at different camera angles.

In Fig. 23, we show more examples of the multi-layered normal maps produced by our P. This figure also demonstrates the effectiveness of P in tackling internal occlusion such as self-occlusion and non-frontal pose of the human subject. Specifically, if we look at the left example in Fig. 23, we notice that the human subject has occluded his face with his left arm (i.e. self-occlusion) in the input image. In order to resolve the self-occlusion, our P generates the 'Back Inner' normal map. This normal map, as shown in the figure, gives us predicted information of the human subject's face. This information will be passed to the Layered-Normals Pixel-aligned Implicit Model, which will use the information to reconstruct the geometry of the human subject's face. If we consider a different scenario where the human subject occludes the right side of his face with his right arm, then the 'Left Inner' normal map will be pivotal for reconstructing the right side of this face (e.g. his right ear). Our P is designed to generate eight different normal maps because there are many different possible body poses that may result in self-occlusion at different view directions, and the eight normal maps ensure robustness to a wide range of body poses. In addition, besides self-occlusion, the eight normal maps also address non-frontal pose problems. Given an input image that show a human subject at a view direction v, we can expect the generated human avatar to appear to be well-constructed when we view it (the avatar) at the view direction v. However, the avatar may not be well-constructed at other view directions, and this is demonstrated by ECON's results in the 1st row of Fig. 2 and the 1st row of Fig. 7. The right and left views of ECON is not well-constructed when ECON is given the front view of a human subject. This is because ECON (Xiu et al., 2023) only predicts and uses the front and back normal maps, and not the left and right normal maps. For this reason, our P generates front, back, left, and right normal maps to ensure that a 360 degree view of our generated avatar will be well-constructed. The effectiveness of this can be observed in our results in the same 1st row of Fig. 2 and the 1st row of Fig. 7.

As aforementioned, SMPL-X normal maps are fed into P. For P to predict the Front Outer, Front Inner, Back Outer, and Back Inner normal maps of a **clothed body**, we provide P with the Front Outer, Front Inner, Back Outer, and Back Inner normal maps of a **SMPL-X mesh** to use as inputs. In other words, there is a correspondence between which SMPL-X normal map is given to P and which clothed body normal map is returned by P. In Fig. 14, we show what would sometimes happen if we did not give P one of the four SMPL-X normal maps. To put it into words, when P is trained without the correct, corresponding SMPL-X normal maps, it will sometimes produce an incorrect normal map as one of its outputs. For example, Fig. 14a shows such a P producing a Front Outer normal map when it is actually supposed to produce a Back Inner normal map instead.

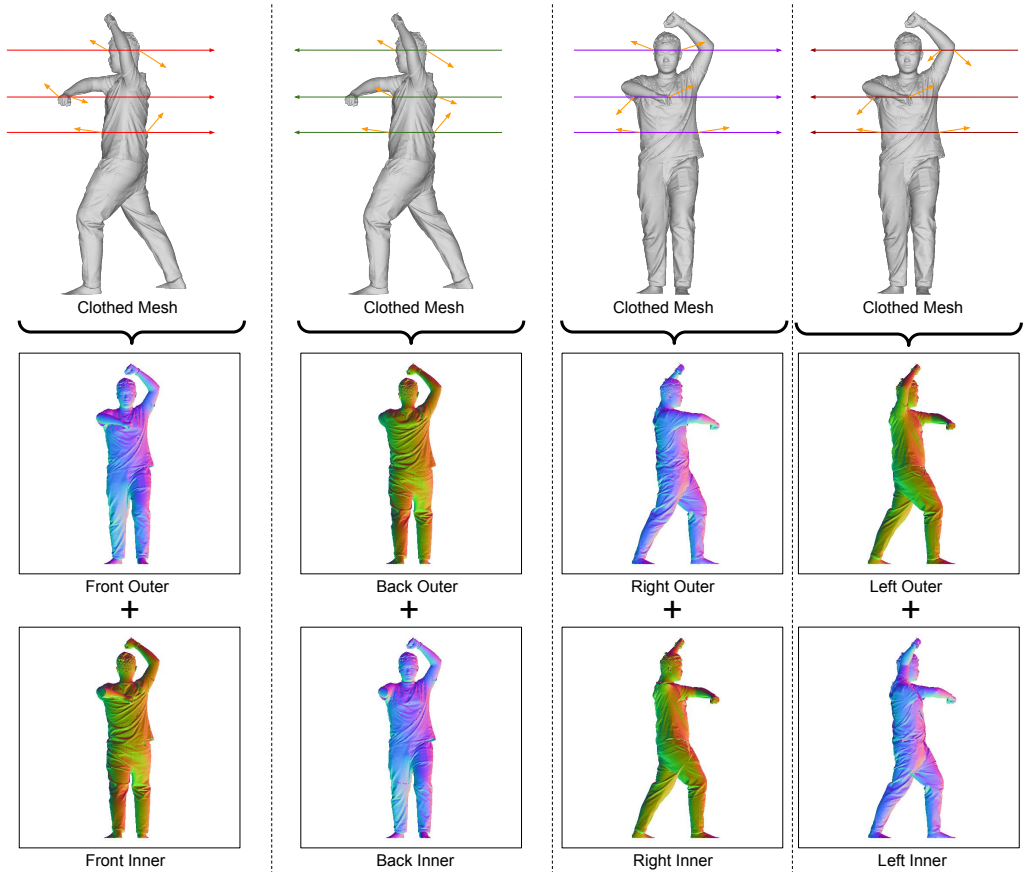


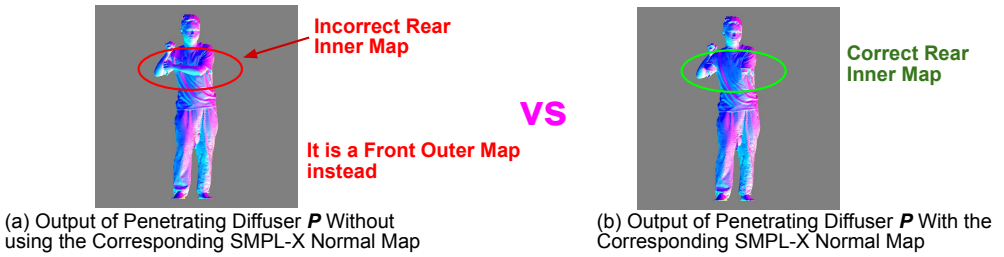
Figure 13: The multi-layered normal maps of a clothed human body (i.e. N_C) consists of 8 normal maps: Front Outer, Front Inner, Back Outer, Back Inner, Right Outer, Right Inner, Left Outer, and Left Inner. Each camera ray (horizontal arrow) is aligned according to a camera direction. Normal vectors are shown as short orange arrows. The first normal vector captured by the ray goes to the 'Outer' map, and the second normal vector captured goes to the 'Inner' map.

A.5 How Multi-dimensional Layered Normals Grid is obtained.

We illustrate how our Multi-dimensional Layered Normals Grid is obtained in Fig. 15.

A.6 More on our Differential Optimizer

In our case, however, we cannot use ICON's differential optimizer directly as we have layered normal maps rather than regular normal maps. Thus, we extend its design and compute the loss between our layered normal maps and the layered surface normals of the clothed human mesh. In addition, the design is also extended such as layered normal maps from different camera angles are allowed to be included during the deformation or adjustment process of the mesh.



(a) Output of Penetrating Diffuser ***P*** Without using the Corresponding SMPL-X Normal Map

(b) Output of Penetrating Diffuser ***P*** With the Corresponding SMPL-X Normal Map

Figure 14: Ablation of our Penetrating Diffusor's use of SMPL-X normal maps

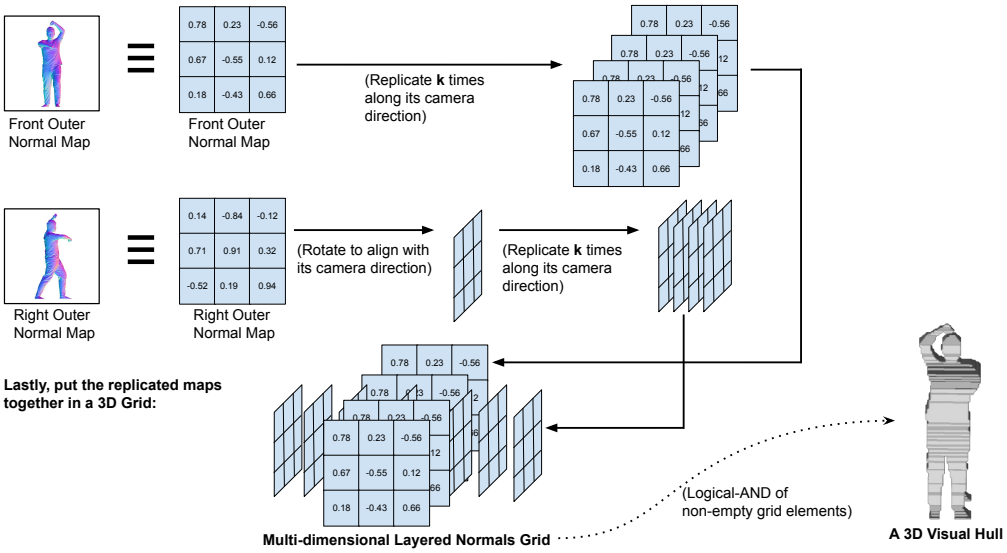


Figure 15: The multi-layered normal maps of a clothed human body (8 maps) are aligned in a 3D grid to form the Multi-dimensional Layered Normals Grid. In the illustration above, we only show how the Front Outer Normal Map and Right Outer Normal Map are used to form the grid, but in reality all 8 maps are used. Only the Front Outer Normal Map (shown above) and the Front Inner Normal Map (not shown) do not need to be rotated to align with their camera direction in the 3D grid because the 3D grid is aligned with the camera direction of these front-facing normal maps. In our work, we use **k** = 128. In addition, each grid element on the Multi-dimensional Layered Normals Grid is a 24-length vector, so the grid has a shape of (24,128,128,128). If we apply a logical-AND operation along the 24 channels to extract the non-empty grid elements (or filter out the empty grid elements), then we obtain a **3D Visual Hull** of a clothed human subject. In practice, we do not manually perform the Logical-AND operation to form a visual hull. Instead, we leave it to the neural network to learn this piece of easily-computable information itself.

A.7 More on our Refining Diffuser (R)

Besides external and internal occlusions, another common problem in in-the-wild images is the low resolution of the given input image. This can occur when the human subject is far away from the camera or when the camera is set to capture more than one human subject. The low-resolution input image may cause existing methods and our Robust-PIFu to produce a clothed human mesh that lacks high-resolution details. For our Robust-PIFu, a low-resolution input image may lead to a clothed human mesh with low-resolution details because the low-resolution input image shares the same resolution as our N_C. For example, if the resolution of the given input image is only 256x256, then the resolution of N_C will also be 256x256. A low-resolution N_C limits the ability of our differential optimizer to refine a clothed human mesh. This is because the differential optimizer directly relies on N_C to guide the refinement process.

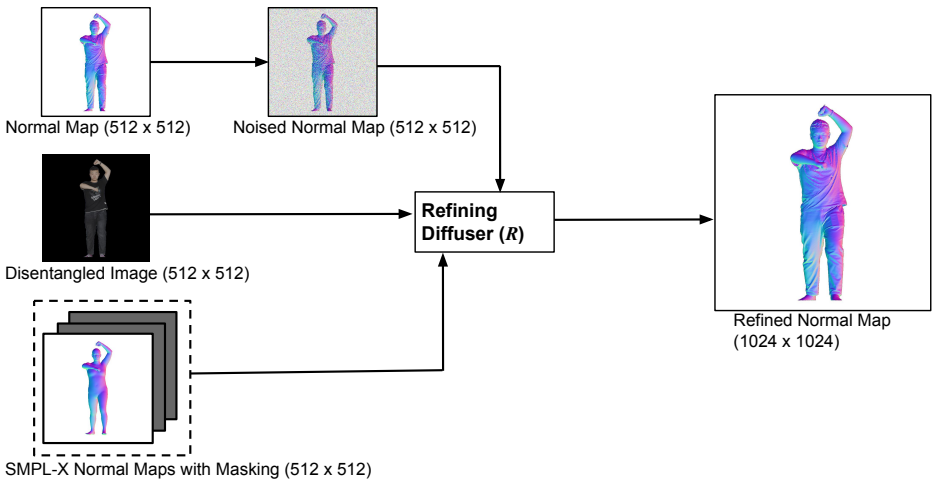


Figure 16: Our Refining Diffuser (R). The 'SMPL-X Normal Map with Masking' is our N_S but 7 out of its 8 maps are masked with zeros. In other words, only 1 of its 8 maps is unmasked, and this 1 unmasked map corresponds to the normal map that will be super-resolved.

To overcome this, we propose our Refining Diffuser (R). The purpose of R is to perform super-resolution on N_C so that our differential optimizer can add high-resolution details to a clothed human mesh despite a low-resolution input image.

To perform normal map super-resolution, which is an under-constrained problem, we again turn to a pretrained large-scale latent diffusion model that has already acquired knowledge from billions of images. Starting from the pretrained model, we inject it with three architectural modifications so that we can include three different conditional priors: 1. A disentangled image 2. A noised normal map 3. SMPL-X normal maps with Masking. Refer to Fig. 16 for an illustration. The disentangled image conditional prior is incorporated using cross-attention (after transforming into CLIP image embedding) and channel concatenation. The other two conditional priors are incorporated via channel concatenation only. With these conditional priors, the task of our R is to denoise a latent embedding and then decode that denoised latent embedding into a high-resolution normal map. R can be used to super-resolve any of the 8 normal maps in our N_C.

The 'noised normal map' conditional prior is obtained by adding Gaussian noise to a normal map from N_C. The reason for adding the noise is because, during testing of R, the normal map from N_C will be a prediction of our P and is thus not perfectly accurate. The noise nudges R away from fully depending on the predicted normal map before producing the super-resolved normal map. This reduces propagation error that may derive from P.

Each run of R will super-resolve 1 normal map from N_C. Since N_C consists of 8 normal maps, we will need to run R eight times to super-resolve the entire N_C.

In each run, we need to indicate to R which of the 8 normal maps should R produce. This is done using the 'SMPL-X normal maps with Masking' conditional prior. 'SMPL-X normal maps with Masking' is our N_S but 7 out of its 8 maps are masked with zeros. In other words, only 1 of its 8 maps is unmasked, and this 1 unmasked map corresponds to the normal map from N_C that will be super-resolved.

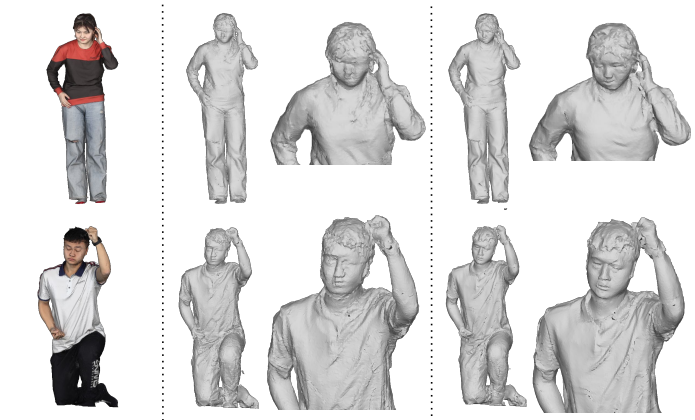
After 8 runs of R, we will obtain a super-resolved N_C, which will be used in our differential optimizer for adding fine, high-resolution details to a clothed human mesh. By doing so, we ensure that our Robust-PIFu is able to produce a clothed human mesh with high-resolution details despite a low-resolution input image.

In summary, a formal formulation of R is represented in the following equation:

$$N_C^S = R(N_C', I_i^d, N_S) \tag{4}$$

where I_i^d is the disentangled image of the i^{th} human subject, and N_C' refers to N_C added with Gaussian noise.

In Fig. 24, we show examples of the high-resolution normal maps produced by our R. This figure also demonstrates the effectiveness of R in super-resolving the normal maps in N_C.



(a) Input Image (b) Without R (c) With R

Figure 17: Ablation of using R and not using it.

Evaluation and Ablation of our Refining Diffuser (R). To evaluate our Refining Diffuser (R), we compare the 3D clothed human meshes produced by our RobustPIFu when R is not used and when R is used in Fig. 17. It is clear that when R is used, the refined clothed human mesh has higher resolution details. In addition, we also conducted an ablation study to investigate the impact of using and not using the noised normal map in R. As shown in Fig. 18, there is a clear benefit to using the noised normal map. The reason for this is that, by noising the input normal map, we give R the freedom to fix any flaws that may be present in the given input normal map. In other words, without noising, if the input normal map is flawed, then R will reproduce and magnify the flaw in higher-resolution. On the other hand, noising the given normal map informs R that there are inherent noises in the input normal map, and the diffuser is required to remove these noises. In the process of doing so, the diffuser actually fixes any flaw in the input normal map as well.

A.8 Results on Internet images from Shutterstock

In Fig. 19, we show the results of our RobustPIFu and ECON on Internet images from Shutterstock. Like all our other qualitative results, this figure shows that our RobustPIFu excels at constructing regions that are obscured/unseen from the given input images. Specifically, unlike ECON, RobustPIFu is able to reconstruct realistic surfaces at the rear/back of the clothed human meshes.

A.9 Results on Image with More than Three Humans and on Images from the MultiHuman and BUFF datasets

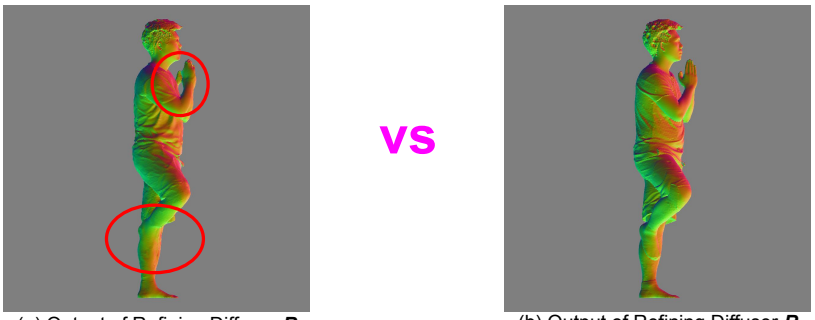
In Fig. 20, we show the results of our RobustPIFu and ECON on a real image with more than three human subjects, a rendered image from the MultiHuman dataset, and a rendered image from the BUFF dataset.

A.10 360 degree view of our results

In place of a video, we show a 360 degree view of our results using a series of images in Fig. 21.

A.11 Why Multi-person and Related Scans from the MultiHuman dataset cannot be used for Evaluation.

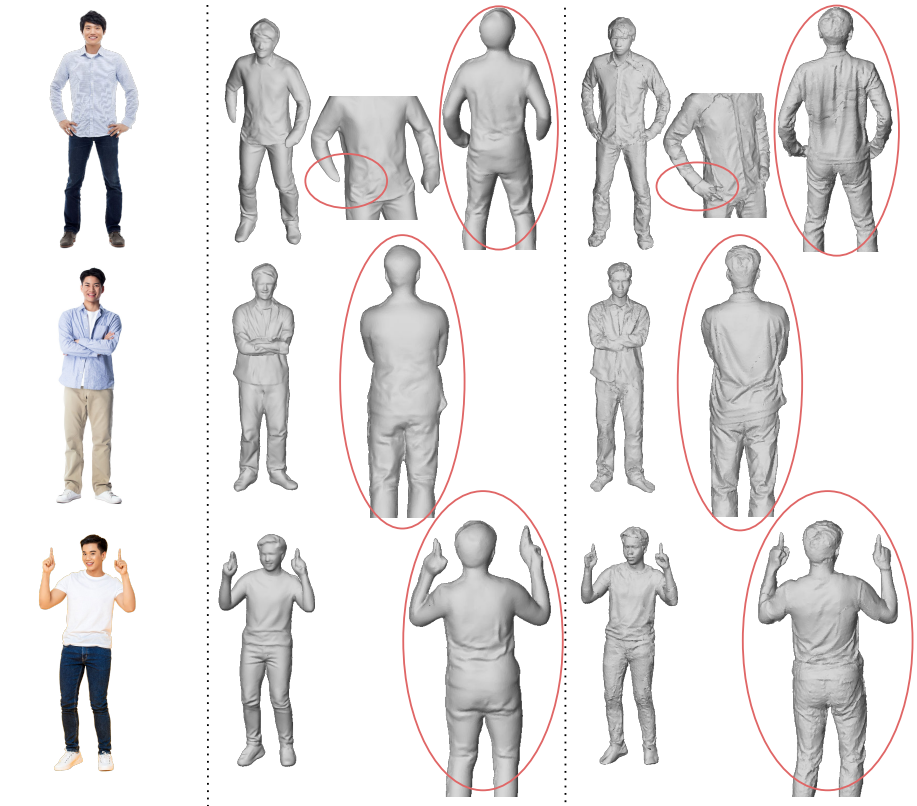
As mentioned earlier, only single human scans from the MultiHuman dataset is used for evaluation. The multi-human scans are excluded. The reason for this is because the multi-human scans contain gaping holes that made them impossible to use for evaluation (the CD and P2S scores will be inaccurate and misleading). This problem is illustrated in Fig. 22. This "gaping hole" problem is also true for 3D scans with human-object interactions in the MultiHuman Dataset. Hence, the only



(a) Output of Refining Diffuser ***R*** Without using Noised Normal Map

(b) Output of Refining Diffuser ***R*** With Noised Normal Map

Figure 18: Ablation of our Refining Diffusor's use of a Noised Normal Map



(a) Input Image (b) ECON (c) Ours

Figure 19: Results on Internet images from Shutterstock.

way we can use the MultiHuman Dataset in our experiment is to use only the 3D scans that have no human-object and human-human interactions, and these are the "single human scans" that we used in our paper.

A.12 Computation Time required by our Robust-PIFu

Short inference time is not a claim made by our paper. However, we acknowledge it is important that our Robust-PIFu does not have a much longer inference time compared to existing models.

Thus, in order to reassure readers, we compare our Robust-PIFu with ECON (Xiu et al., 2023) in terms of the average time required to produce a clothed human mesh during testing. From this experiment, we observe that Robust-PIFu takes 126.42 seconds, while ECON takes 138.11 seconds. The 126.42 seconds include the time required for our D and P generate their outputs, and the time

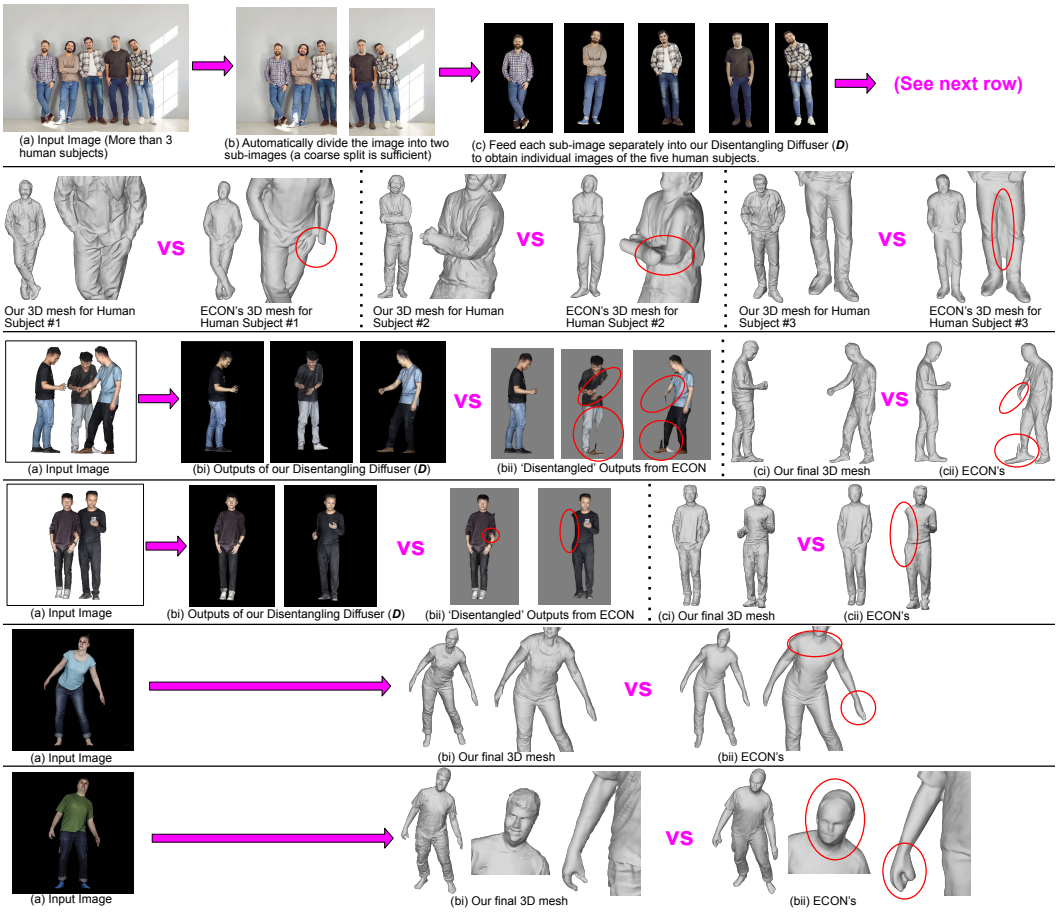


Figure 20: Additional Results. The first two rows illustrate how a real image (from Adobe Stock) with more than three humans will be handled by our RobustPIFu. The third and fourth rows show results from the MultiHuman dataset, and the fifth and sixth rows show results from the BUFF dataset.

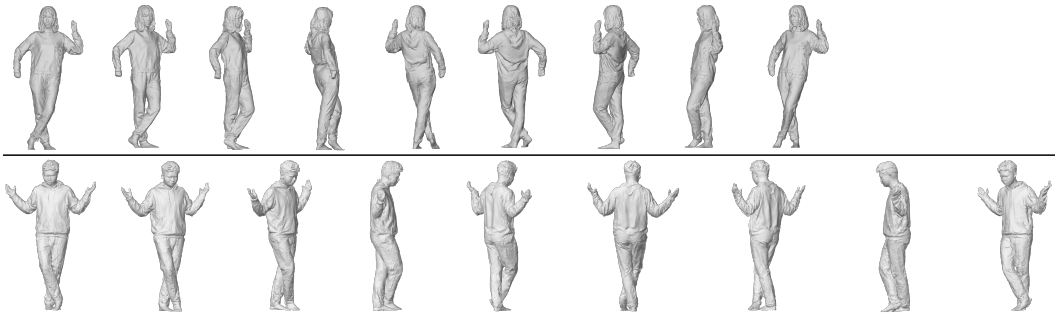


Figure 21: 360 degree view of results from our RobustPIFu. Results using two different input images are shown.

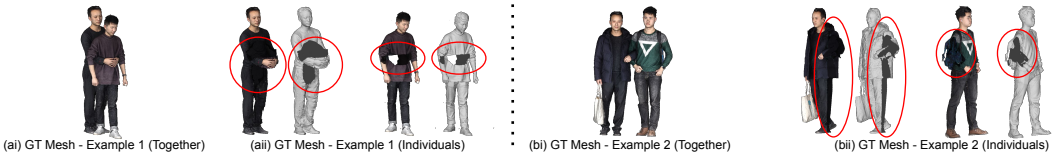


Figure 22: Gaping holes observed in the multi-human scans from the MultiHuman dataset. These holes made it infeasible to use multi-human scans from the dataset for evaluation purposes.

Table 3: Additional Ablation study on Layered-Normals Pixel-aligned Implicit Model.

Methods	THuman2.0	
	CD (10^{-5})	P2S (10^{-5})
Robust-PIFu w/o Multi-D Grid (No N_S)	8.643	8.227
Robust-PIFu w/o Multi-D Grid (With N_S)	8.551	8.132

required to construct the Multi-dimensional Layered Normals Grid. During the experiment, the same hardware (1 NVIDIA RTX A5000 GPU) is used to test both Robust-PIFu and ECON.

A.13 Other Limitations

Our proposed method uses generative models, which may result in hallucinations (e.g. predict clothes wrinkles that are not present at the back of a human subject). We show a few examples of this in Sect. A.19. In addition, we did not use the RenderHuman dataset for training or evaluation because it is a commercial dataset that is extremely costly to access and thus unavailable to us.

A.14 Implementation Details

Our D, P, and R are initialized using the same pretrained latent diffusion model that was employed in Zero-1-to-3 (Liu et al., 2023b). This latent diffusion model is pretrained on the LAION datasets (Schuhmann et al., 2022).

The pretrained latent diffusion model is finetuned with THuman2.0 images, and this is done with an AdamW (Loshchilov & Hutter, 2017) optimizer that has a learning rate of 10^{-4}. The batch size used is 72. The training images have a resolution of 512x512, and we use a latent dimension of 64x64x4. This means that our D aims to denoise a 64x64x4 noised encoding that represents an image of a human subject. In addition, this also implies that the input image I has to be encoded into a 64x64x4 encoding. Also, the denoised encoding generated by the denoising U-Net will be decoded into a 512x512 'disentangled' image.

During inference, we use 200 DDIM sampling steps.

As for our Layered-Normals Pixel-aligned Implicit Model, it is trained using a RMSprop optimizer that has an initial learning rate of 10^{-3}. Its encoder is a stacked hourglass network (Newell et al., 2016) with 4 stacks. Its MLP has layers of the following dimensions: (257+64+24, 1024, 512, 256, 128, 1). The 3rd, 4th, and 5th layers have skip connections. The '+64' is for the features generated by the 3D-CNN that processes the Multi-dimensional Layered Normals Grid, and the '+24' is for the skip connection that directly feeds values from the Multi-dimensional Layered Normals Grid into the MLP. The design of the MLP follows the original PIFu model (Saito et al., 2019), except that its input dimension has been increased by a total of 88 to accommodate the features generated by the 3D-CNN and the values fed by the skip connection.

Every component of our Robust-PIFu is finetuned or trained using NVIDIA RTX A5000 GPUs. More implementation details can be found in our source code, which will be made publicly available.

A.15 Additional Ablation study on Layered-Normals Pixel-aligned Implicit Model

In this subsection, we discuss the inclusion of N_S in our Layered-Normals Pixel-aligned Implicit Model. We observe in Tab. 3 that it was beneficial to include N_S instead of excluding it. We believe that it was due to the fact that N_S does not contain pixels that pertain to clothes or hair. The combination of N_S and N_C thus gives our pixel-aligned implicit model information of which pixels belong to clothes and hair. Clothes and hairs in a 3D mesh are often thin rather than thick, and this information is helpful for guiding the pixel-aligned implicit model in constructing the 3D mesh.

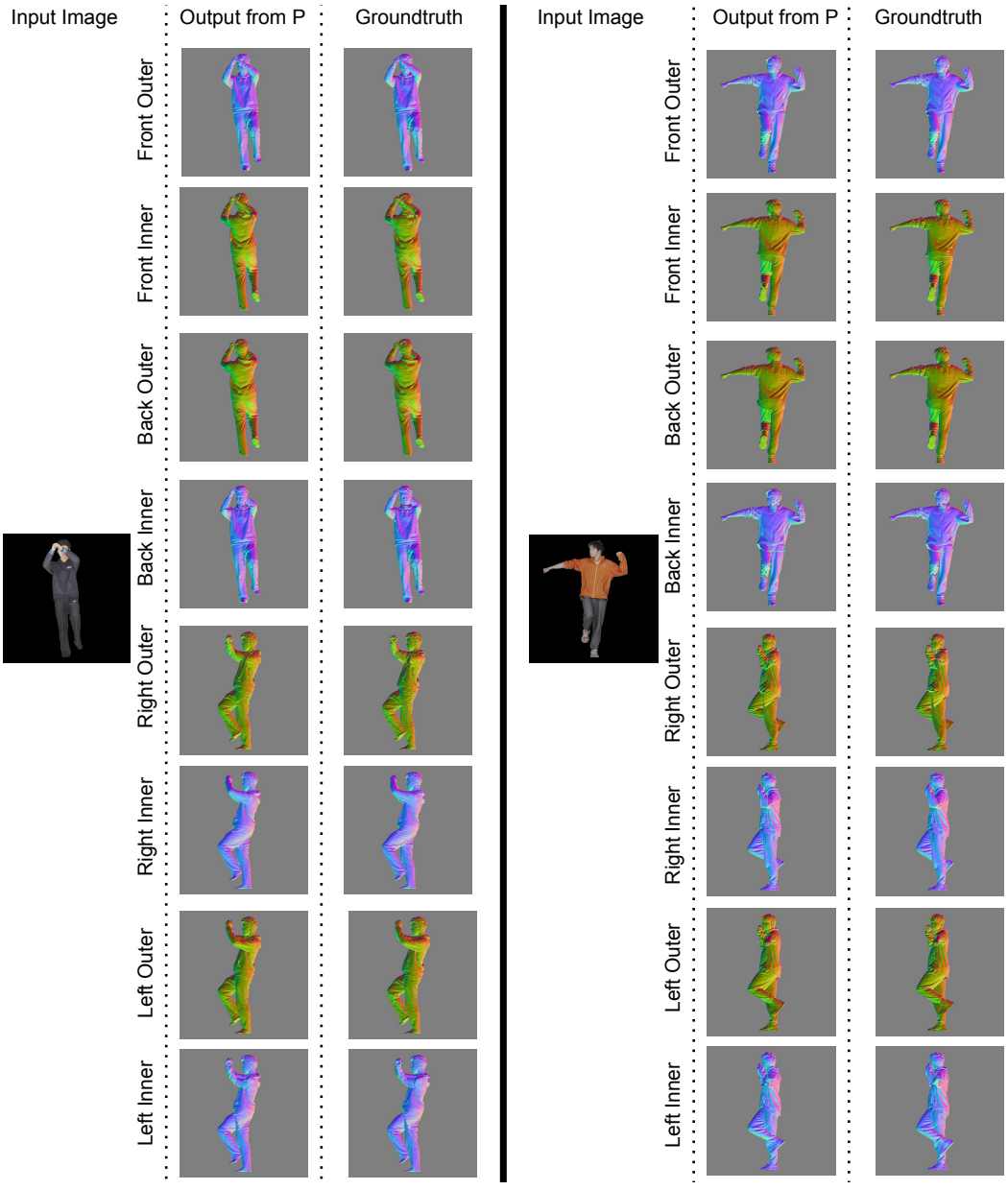


Figure 23: Additional Results on the Outputs produced by our P

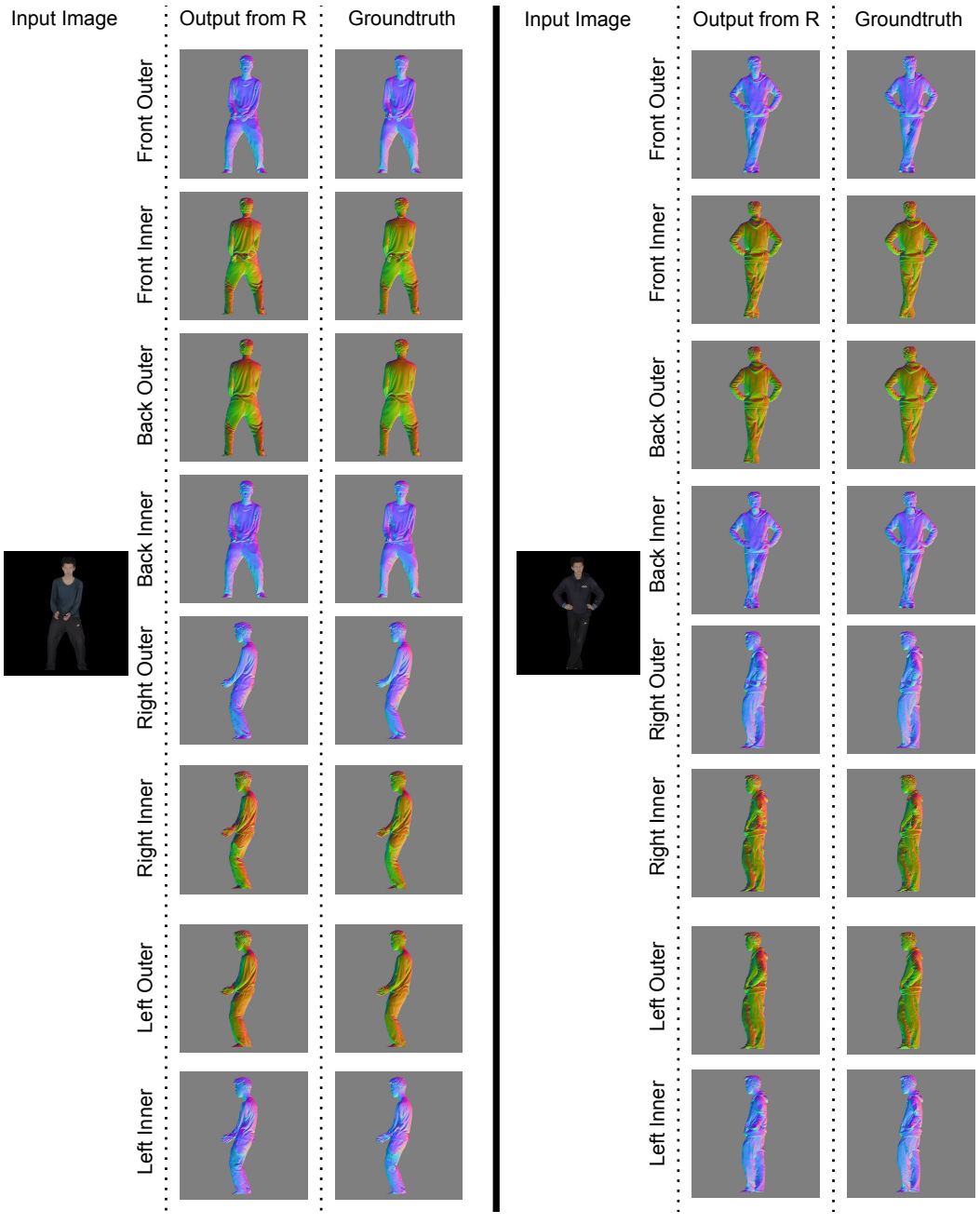


Figure 24: Additional Results on the Outputs produced by our R. The input images shown above are 512x512 in resolution while the normal maps ('Output from R' and 'Groundtruth') are 1024x1024 in resolution.

A.16 ADDITIONAL RESULTS #1

(a) Input Image (b) Our Results

Figure 25: Results pertaining to the 4th row of Fig. 7, but on the second human subject instead.

A.17 ADDITIONAL RESULTS #2

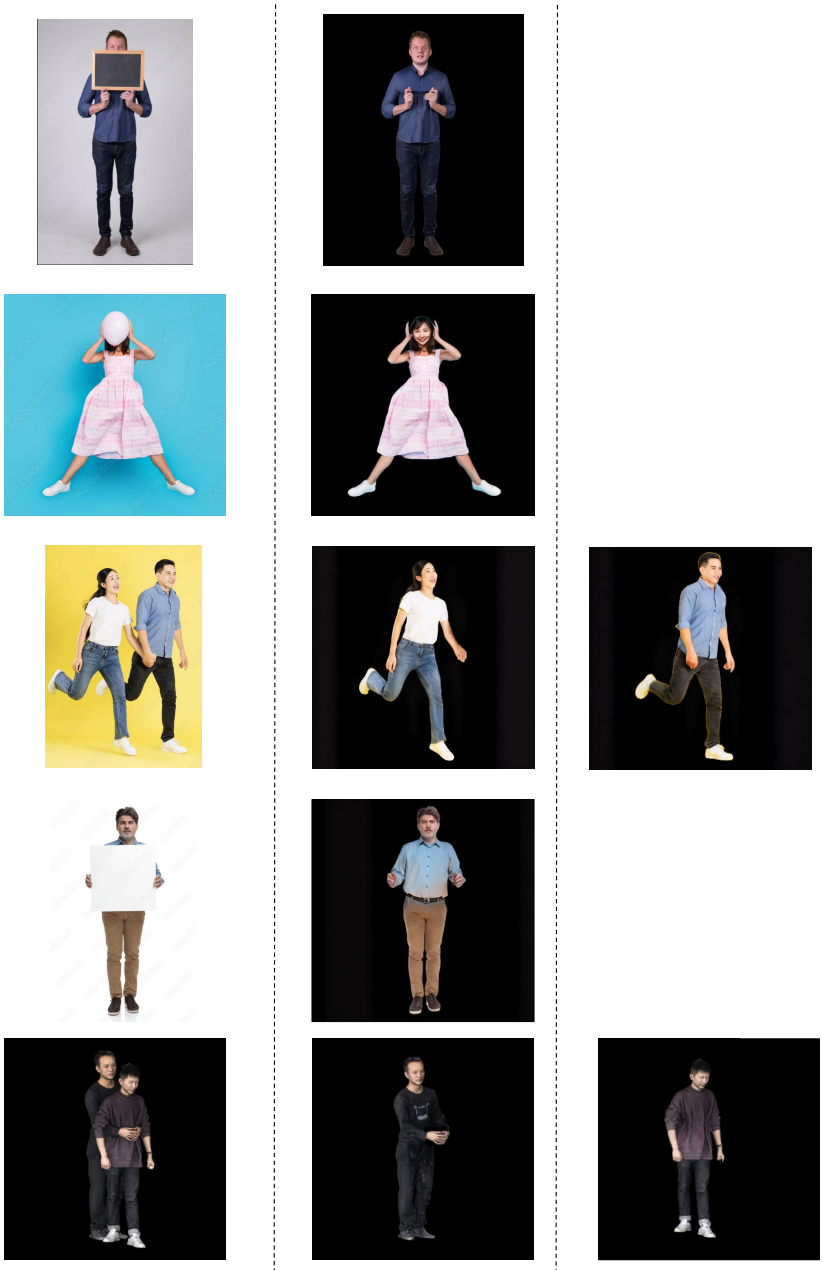


Figure 26: Additional Results pertaining to different degree of occlusion.

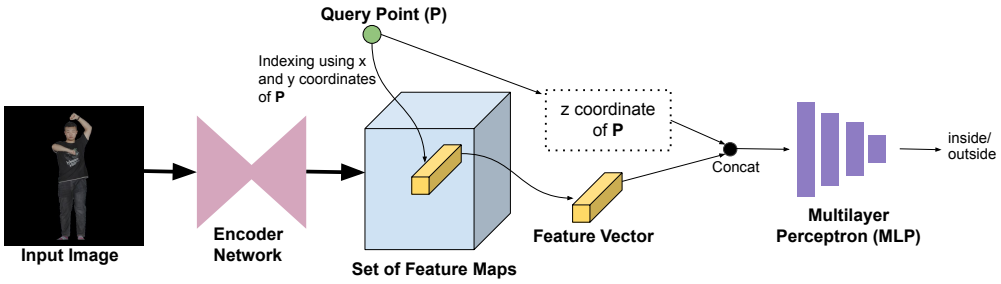


Figure 27: Pipeline and Architecture of a generic Pixel-aligned implicit model. An input image is processed by an encoder network into a set of feature maps. Query points are used to index the set of feature maps to retrieve feature vectors. The feature vectors are then fed into a MLP that will predict the occupancy of the query points.

A.18 Detailed Explanations on Related Works and Concepts

A.18.1 Detailed Explanations on Pixel-aligned Implicit Models

In this section, we will explain how a vanilla pixel-aligned implicit model works. The concepts explained in here are derived from the PIFu paper (Saito et al., 2019), which is one of the earliest works on pixel-aligned implicit model.

In Fig. 27, we show the pipeline of a generic pixel-aligned implicit model. At the start of the pipeline, an input image is fed into an encoder network, which can be any 2D Convolutional Neural Network. Usually, a stacked hourglass network is used as the encoder network. The encoder outputs a set of feature maps. The set of feature maps, as a whole, has the shape of (C, H, W), where C is the number of channels, H is the height of a feature map, and W is the width of a feature map.

Spatially, we say that the set of feature maps is pixel-aligned with the input image because each pixel on the input image can be spatially mapped to a feature vector on the set of feature maps. In other words, the set of feature maps corresponds to the 3D camera space of the input image.

To train the pixel-aligned implicit model, 3D points are sampled within the 3D camera space of the input image, and these 3D points are called query points. For each query point, its x and y coordinates are used to index the set of feature maps. After indexing, each query point will retrieve a corresponding feature vector. The z coordinate of the query point is then concatenated with that feature vector before being used as input in a Multilayer Perceptron (MLP).

Finally, the MLP will output the estimated occupancy of that query point. The occupancy of a query point refers to a binary prediction of whether the query point is located inside or outside a groundtruth 3D human mesh that has been transformed into the camera space of the input image. In practice, the occupancy predicted by the MLP is a continuous value that has a range between 0 and 1. The value of 0.5 is interpreted as where the mesh surface is located at. Also, any value between 0.5 and 1 is interpreted as being inside the mesh, and any value between 0 and 0.5 is interpreted as being outside of the mesh.

During inference, a 3D grid of query points will be fed into the pixel-aligned implicit model's pipeline, and a 3D grid of occupany predictions will be obtained. Then, the Marching Cubes algorithm Lorensen & Cline (1987) will be applied on the 3D grid of occupancy predictions in order to extract a predicted 3D human mesh.

A.18.2 Detailed Explanations on the Zero-1-to-3 model

Zero-1-to-3 (Liu et al., 2023b) is a work that proposes the use of large-scale diffusion models in the problem of Single-view 3D Object Reconstruction. In their work, the authors of Zero-1-to-3 demonstrated that a large, pre-trained latent diffusion model can be manipulated to produce a novel view of an object from any specified viewpoint. This can be done by finetuning the pretrained model with an input RGB view of the object, a relative camera transformation, and a transformed RGB view of the same object. After being finetuned, the model, when provided with an input view of the object, can generate numerous novel views of that object. These new views are then utilized to reconstruct a 3D mesh of the object using Score Jacobian Chaining (SJC) (Wang et al., 2023a).

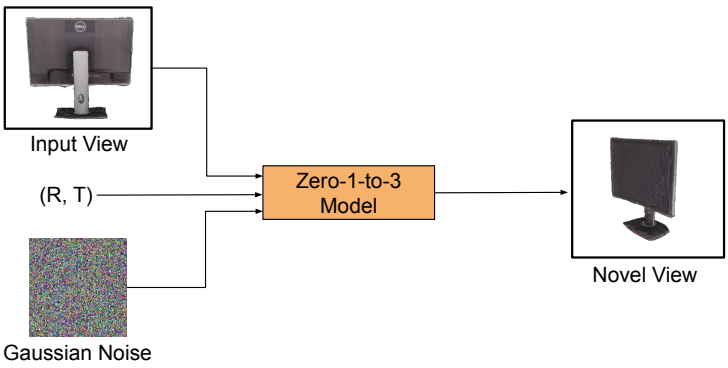


Figure 28: Pipeline of Zero-1-to-3. An input view and a camera extrinsics (R,T) are used by the Zero-1-to-3 model to predict a novel view of an object. The Gaussian noise is required as Zero-1-to-3 uses a latent diffusion model.

The Zero-1-to-3 model can be condensed into the following equation:

$$\tilde{x}_{R,T} = f(x, R, T), \quad R \in \mathbb{R}^{3\times3}\ , \quad T \in \mathbb{R}^3 \tag{5}$$

Where $\tilde{x}_{R,T}$ denotes the synthesized image at a novel viewpoint, x denotes the given input image. R and T refers respectively to the relative camera rotation and translation between x and $\tilde{x}_{R,T}$. The aim of Zero-1-to-3 is to predict a $\tilde{x}_{R,T}$ that is perceptually similar to the groundtruth novel view $x_{R,T}$. In addition, we provided an illustration of Zero-1-to-3 in Fig. 28.

To train Zero-1-to-3, the authors prepared a dataset of paired images and their relative camera extrinsics i.e. $\{(x, x_{R,T}, R, T)\}$. Zero-1-to-3 used a latent diffusion architecture with an encoder (denoted as $\mathcal{E}$), a denoising U-Net (denoted as ϵ_θ), and a decoder. The diffusion time step t ranges from 1 to 1000. Zero-1-to-3 aims to solve the following objective function during the finetuning process:

$$\min_\theta \mathbb{E}_{z\sim\mathcal{E}(x),t,\epsilon\sim N(0,1)}||\epsilon - \epsilon_\theta(z_t, t, c(x, R, T))||_2^2 \tag{6}$$

After the finetuning process, Zero-1-to-3 (or f) will generate an image by performing iterative denoising from a Gaussian noise image by using the conditional prior $c(x, R, T)$, which is an embedding of the input image and the relative camera extrinsics.

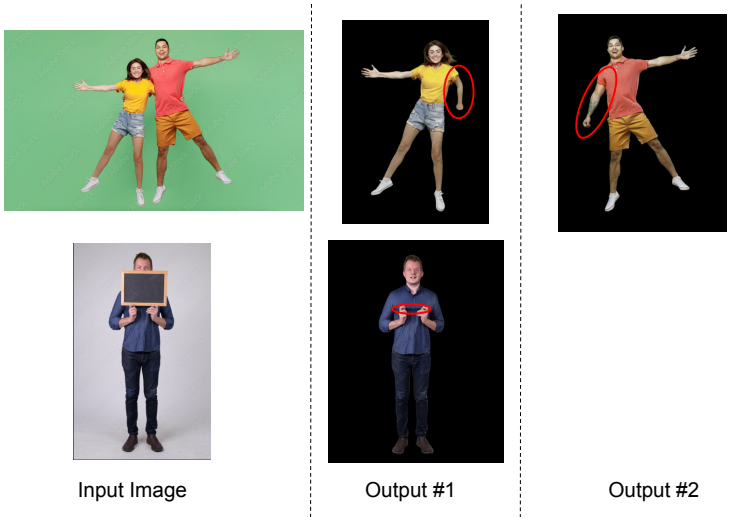


Figure 29: Analysis of Failure Cases. The red circled regions indicate the erroneous areas.

A.19 Analysis of Failure Cases

In the first row of Fig. 29, we observe that the woman's occluded left arm is being generatively filled up by our Disentangling Diffuser (D). But if we inspect the input image closely, we can see that fingers from the woman's left hand are actually grabbing the man's waist. Thus, the left arm predicted by our D has an incorrect position i.e. the left arm should have been straightened, and the fingers should be spread out.

In addition, in the second row of Fig. 29, there is a line of shadow in the output generated by our D. This shadow is hallucinated and an error of D.

These errors are due to insufficient training data. In-the-wild test images have challenging body poses and novel objects that cause occlusion. Consequently, the training data used must be large and diverse enough so as to ensure D can cope with such test images. However, such training data can be challenging to obtain because there is a lack of existing datasets that address such occlusion problems.

A.20 Future Works

For our future works, we intend to address the failure cases that we raised above by building a large-scale dataset that addresses external and internal occlusion problems. We believe that such a dataset will allow us to train ours and related models more effectively and be an important contribution to this field.

Moreover, we will also be exploring if the occlusion removal can be given a finer degree of control. Specifically, rather than just ensuring that the occlusion is removed, we want to specify how it will be removed. For example, if there is an occluded hand in the input image, we wish to control the hand pose of the hand that we generated to replace the occluded hand.

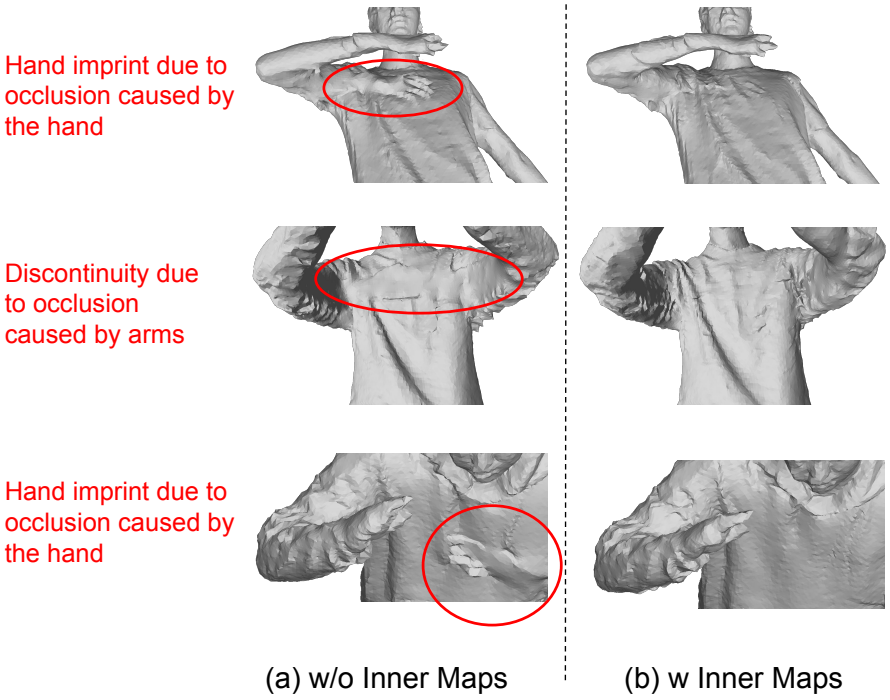


Figure 30: Ablation with and without Inner maps.

A.21 Ablation with and without Inner maps

In Fig. 30, we demonstrate the impact of using the inner maps generated by our Penetrating Diffuser. As shown in the figure, if inner maps are not used, the geometry of occluded regions caused by self-occlusion will contain erroneous and noisy artifacts.