R-Cyclic Diffuser: Reductive and Cyclic Latent Diffusion for 3D Clothed Human Digitalization

Kennard Yanting Chan
Nanyang Technological University, Singapore
Institute for Infocomm Research, A*STAR
kenn0042@e.ntu.edu.sg

Fayao Liu
Institute for Infocomm Research, A*STAR
liu_fayao@i2r.a-star.edu.sg

Guosheng Lin
Nanyang Technological University, Singapore
gslin@ntu.edu.sg

Chuan Sheng Foo
Centre for Frontier AI Research, A*STAR
Institute for Infocomm Research, A*STAR
foo_chuan_sheng@i2r.a-star.edu.sg

Weisi Lin
Nanyang Technological University, Singapore
wslin@ntu.edu.sg

Abstract

Recently, the authors of Zero-1-to-3 demonstrated that a latent diffusion model, pretrained with Internet-scale data, can not only address the single-view 3D object reconstruction task but can even attain SOTA results in it. However, when applied to the task of single-view 3D clothed human reconstruction, Zero-1-to-3 (and related models) are unable to compete with the corresponding SOTA methods in this field despite being trained on clothed human data.

In this work, we aim to tailor Zero-1-to-3's approach to the single-view 3D clothed human reconstruction task in a much more principled and structured manner. To this end, we propose R-Cyclic Diffuser, a framework that adapts Zero-1-to-3's novel approach to clothed human data by fusing it with a pixel-aligned implicit model.

R-Cyclic Diffuser offers a total of three new contributions. The first and primary contribution is R-Cyclic Diffuser's cyclical conditioning mechanism for novel view synthesis. This mechanism directly addresses the view inconsistency problem faced by Zero-1-to-3 and related models. Secondly, we further enhance this mechanism with two key features - Lateral Inversion Constraint and Cyclic Noise Selection. Both features are designed to regularize and restrict the randomness of outputs generated by a latent diffusion model. Thirdly, we show how SMPL-X body priors can be incorporated in a latent diffusion model such that novel views of clothed human bodies can be generated much more accurately. Our experiments show that R-Cyclic Diffuser is able to outperform current SOTA methods in single-view 3D clothed human reconstruction both qualitatively and quantitatively. Our code is made publicly available at https://github.com/kcyt/r-cyclic-diffuser.

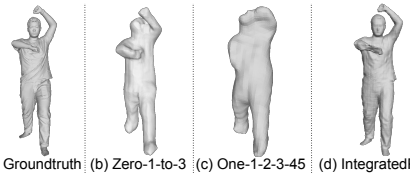


Figure 1. Results of Zero-1-to-3 [7], One-2-3-45 [6], and IntegratedPIFu [2] when trained and tested on clothed human subjects.

1. Introduction

Recently, Zero-1-to-3 [7] demonstrated that a latent diffusion model, which has been pre-trained on Internet-scale image data, can be controlled to synthesize a novel view of a subject of interest from any specified viewpoint. To do so, the pretrained model has to be finetuned with an input RGB view of a subject, a relative camera transformation, and a transformed RGB view of the same subject. Once finetuned, the model, when given an input view of a subject, will be able to generate a large number of novel views of that subject. These novel views are then used to reconstruct a 3D mesh of that subject using Score Jacobian Chaining (SJC) [19]. Zero-1-to-3 showed that it is able to outperform SOTA results in 3D object reconstruction. This is significant because Zero-1-to-3 shows how the massive internet image data can be leveraged on to solve a 3D reconstruction task.

One-2-3-45 [6] builds on Zero-1-to-3 by estimating the

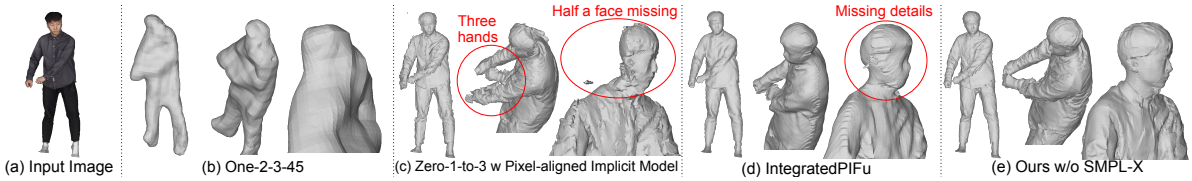


Figure 2. Our R-Cyclic Diffuser compared with SOTA models when SMPL-X priors are unavailable.

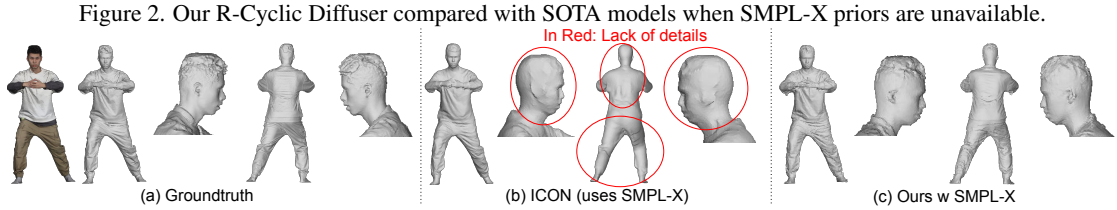


Figure 3. Our R-Cyclic Diffuser compared with SOTA models when SMPL-X priors are available.

elevation angle of the input image and replacing the SJC with SparseNeuS [8]. Compared to Zero-1-to-3, One-2-3-45 showed that it can generate a more accurate 3D mesh in a shorter amount of time. One-2-3-45, though, continues to rely on the novel views generated by a frozen Zero-1-to-3.

This prompts the question of whether we can extend the Zero-1-to-3 and One-2-3-45 models to tackle the single-view 3D clothed human reconstruction task too. Experimental results (see Fig. 1) show that while a direct application of Zero-1-to-3 and One-2-3-45 to human data will allow us to reconstruct 3D clothed human meshes, their results fall significantly short of competing with the current SOTA methods (e.g. IntegratedPIFu [2]) in this field. Hence, we aim to tailor Zero-1-to-3 to this field in a much more principled and structured manner. Specifically, we do so by fusing Zero-1-to-3's novel approach with a pixel-aligned implicit model, which is a class of methods [2, 15, 16, 20] that has consistently achieved SOTA results in the field. In doing so, we hope to leverage Internet-scale data to compete with and even surpass SOTA methods.

Our proposed model, called R-Cyclic Diffuser, is a latent diffusion model that first generates a sparse set of views in a sequential manner. These views are then fused together using a multi-view pixel aligned implicit model [15]. R-Cyclic Diffuser offers a total of three new contributions.

The first and primary contribution is R-Cyclic Diffuser's cyclical conditioning mechanism for novel view synthesis. This mechanism resolves the view inconsistency problem (see later) that plagued Zero-1-to-3.

The second contribution of R-Cyclic Diffuser is the introduction of Lateral Inversion Constraint and Cyclic Noise Selection, which are key features designed to regularize and restrict the randomness of outputs generated by our latent diffusion model. This will be elaborated later.

Lastly, the third contribution of R-Cyclic Diffuser proposes how SMPL-X body priors [10] can be incorporated in a latent diffusion model such that that novel views of clothed human bodies can be generated much more accurately. This is important because Zero-1-to-3 and related models are designed for 3D object reconstruction and do not provide any design consideration for incorporating priors that are specific to clothed human reconstruction.

2. Related Work

2.1. Single-view Object Reconstruction

Recently, large-scale 2D diffusion models (e.g., DALL-E [12], Imagen [14], and Stable Diffusion [13]) have demonstrated their ability to learn a wide range of visual concepts from Internet-scale image datasets. This has led a growing number of researchers to leverage on the models' extensive priors about our 3D world and utilize them in 3D generative tasks (e.g. DreamFusion [11], and Magic3D [5]). They typically start with a 3D NeRF representation and optimize it by first generating 2D images at different viewpoints using differentiable rendering. The images are then fed to a pre-trained 2D diffusion model that would compute a loss function that guides the 3D shape optimization.

A very recent work, Zero-1-to-3 [7], explores a new approach to that by starting with the pre-trained diffusion model instead. The pre-trained diffusion model is finetuned such that it learns how to generate a novel view of an object at a desired viewpoint, given only an input image of that object and a specification of that desired viewpoint. Once finetuned, the diffusion model is used to generate multiple novel views of the object, and these views are used to optimize a 3D NeRF representation. With that, Zero-1-to-3 attained SOTA results on single-view 3D reconstruction.

Shortly after, the authors of One-2-3-45 [6] build on Zero-1-to-3 by adding a module that estimates camera elevation and a SparseNeuS [8]. One-2-3-45 showed that it outperformed Zero-1-to-3 in terms of speed and accuracy. However, as we showed in Fig. 1, neither Zero-1-to-3 nor One-2-3-45 are able to outperform the SOTA methods in single-view clothed human reconstruction.

2.2. Single-view Human Reconstruction

In 3D reconstruction of clothed humans, a class of models that have consistently achieved the SOTA results is the pixel-aligned implicit models. These models learn an im-

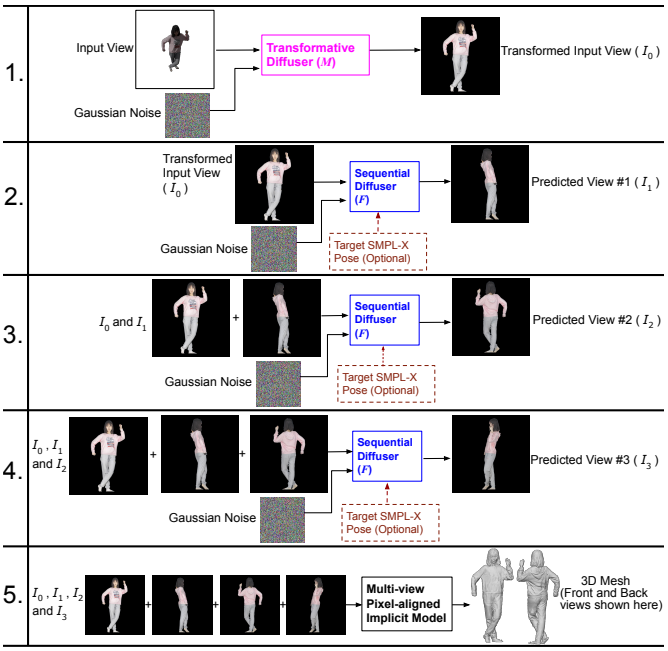


Figure 4. Overview of R-Cyclic Diffuser.

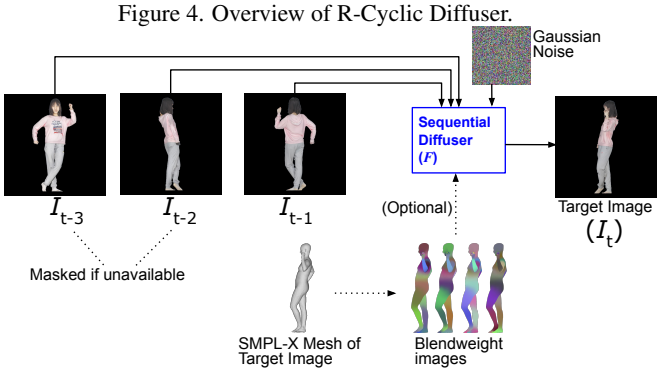


Figure 5. Specifics of our F. Use of SMPL-X Mesh is optional.

plicit function that represents the surface of a clothed human body. From the learned implicit function, a mesh of a human body can be extracted using the Marching Cubes algorithm [9]. Pixel-aligned implicit models include PIFu [15], ICON [20], IntegratedPIFu [2], FSS [3] and more.

Unlike Zero-1-to-3 and One-2-3-45, pixel-aligned implicit models have not fully explored the potential of leveraging a latent diffusion model pretrained with Internet-scale image data to improve their own 3D reconstruction capabilities. Thus, in order to address this research gap, our work aims to fuse the novel approach coined by Zero-1-to-3 with a pixel-aligned implicit model for the task of single-view clothed human reconstruction.

3. Method

Our R-Cyclic Diffuser consists of two latent diffusion models (M and F), as shown in Fig. 4. Both latent diffusion models are pretrained with Internet-scale image data (as done in Zero-1-to-3 [7]) and subsequently finetuned by us. The first latent diffusion model (i.e. Transformative Diffuser or M) will take an image and a Gaussian noise sample as inputs and transform that image into an 'ideal' image. An 'ideal' image is an image that appears to be captured under ideal conditions such as perfect lighting, camera angle (elevation but not azimuth), focal length etc. The ideal image (I_0) serves as a strong and clear reference point for the subsequent views that we are about to generate.

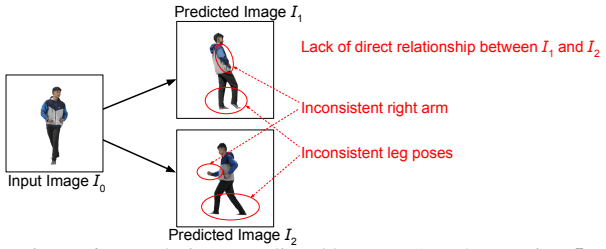


Figure 6. Novel views predicted by Zero-1-to-3 [7] using I_0

We will then apply the second latent diffusion model (i.e. Sequential Diffuser or F). F will first be given I_0 and a Gaussian noise sample as inputs. F will then output a novel view I_1, which is a θ degree azimuth rotation of the viewpoint in I_0. In our work, we set θ to be 90°. The novel view predicted by F will be an 'ideal' image as well.

Next, F will use I_0, I_1, and a Gaussian noise sample to generate I_2, which is a 90° azimuth rotation of I_1. Likewise, I_2 will be an 'ideal' image too. Lastly, I_0, I_1, I_2, and a Gaussian noise sample will be used by F to produce I_3. In total, we obtain 4 images (I_0, I_1, I_2, and I_3).

If we apply F on I_1, I_2, I_3, and a Gaussian noise sample, then F will return a predicted view that is at the same viewpoint as I_0. This is deliberate and is a hallmark of our cyclic conditioning mechanism. What this creates is a geometric relationship between the different views, and we will show how we exploit this later.

Optionally, if a SMPL-X mesh [10] is available as an input, we can also use it as an additional input to F (see Fig. 5). This will be elaborated later.

Finally, I_0, I_1, I_2, and I_3 will be fed into a multi-view pixel-aligned implicit model [15] that will output a 3D clothed human mesh (see bottom of Fig. 4). The specific design of the multi-view pixel-aligned implicit model used by us is already established in existing works. Thus, we will describe its exact implementation in our Supp. Mat.

Now, we will elaborate on the three contributions of R-Cyclic Diffuser: 1. Cyclic Conditioning Mechanism, 2. Lateral Inversion Constraint and Cyclic Noise Selection and 3. Incorporating SMPL-X Blendweight Priors.

3.1. Cyclic Conditioning Mechanism

View inconsistency is the problem where the predicted images do not reconcile with one another (e.g. a man has a certain pose in image 1 but has a different pose in image 2.). This problem is clearly an issue for Zero-1-to-3 in Fig. 6, where Zero-1-to-3 assumes a certain leg pose in I_1, but

then it assumes another leg pose in I_2, causing view inconsistency between I_1 and I_2.

Our Cyclical Conditioning Mechanism addresses the view inconsistency problem by modifying the Zero-1-to-3's prediction function from Eqn. 1 to Eqn. 2.

$$I_t = F(I_0, R, T) \tag{1}$$

where I_t is the predicted image, F is a latent diffusion model, I_0 is the input image, and (R, T) is the relative camera rotation and translation between I_0 and I_t.

$$I_t = F(I_{t-1}, I_{t-2}, I_{t-3}), \quad I_t, I_{t-1}, I_{t-2}, I_{t-3} \in \mathbb{S} \tag{2}$$

where I_t, I_{t-1}, I_{t-2}, and I_{t-3} are the predicted images at time t, $t-1$, $t-2$, and $t-3$ respectively. $\mathbb{S}$ is a set of images captured in 'ideal' conditions (to be elaborated later). Also, for the first prediction, $I_0 = I_{t-1}$ while I_{t-2} and I_{t-3} are masked with zeros.

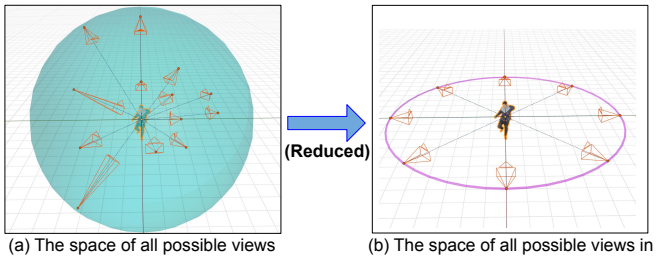
The first problem observed in Eqn. 1 is that Zero-1-to-3 does not take previous predictions into account when predicting I_t, contributing to the view inconsistency problem. To resolve this, our Eqn. 2 is designed such that every prediction is conditioned on the previous predictions. This allows I_t to reconcile with $\{I_{t-1}, I_{t-2}, I_{t-3}\}$ and significantly reduces the view inconsistency problem.

The second problem observed in Eqn. 1 is that Zero-1-to-3, via the use of (R, T), allows the predicted novel view to be at almost any viewpoint. In other words, the output space of F in Eqn. 1 is the teal sphere in Fig. 7a. Zero-1-to-3 needs to do this because it requires images to be from very diverse viewpoints when it optimizes its NeRF representation. But as R-Cyclic Diffuser is using a pixel-aligned implicit model, such a diverse set of viewpoints is excessive and unnecessary. Hence, by setting (R, T) to a constant and introducing $\mathbb{S}$ in Eqn. 2, we drastically reduced the output space of F from the teal sphere in Fig. 7a to the magenta ring in Fig. 7b. Why this is so will be explained later. In the end, our F has a simplified output space and is less susceptible to prediction error. This mitigates any view inconsistency that is caused by F not predicting images accurately.

Our Cyclic Conditioning Mechanism consists of two parts: 1. Ideal Subset Transformation (deals with the second problem) 2. Sequential Conditioning (deals with the first problem).

3.1.1 Ideal Subset Transformation (IST)

In most applications, the input image that would be given to Zero-1-to-3 (or related models) is often an in-the-wild image. Thus, it is likely that the image is captured under imperfect camera position, calibration, or lighting (see Fig. 9). As such, the image itself may contain ambiguous or utterly limited information, forcing Zero-1-to-3 to guess what



(a) The space of all possible views in Zero-1-to-3 (Teal sphere)

(b) The space of all possible views in our R-Cyclic Diffuser (Magenta ring)

Figure 7. A Zero-1-to-3's output can be anywhere **within** the teal sphere, but a R-Cyclic Diffuser's output will be a point **on** the magenta ring. Be informed that the input domain of R-Cyclic Diffuser is restricted by $\mathbb{S}$ in Eqn. 2.

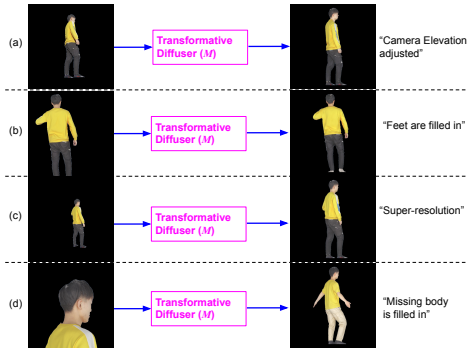


Figure 8. Transforming 'in-the-wild' images into 'ideal' images. Every image here has the same resolution of 512×512.

is really captured by the image. Zero-1-to-3 uses this input image to generate other images, but this input image, because of its ambiguity, serves as a poor reference point for the generated images, resulting in view inconsistency among the generated images. It would be more sensible for the ambiguity in the input image to be cleared up, even if it inevitably involves putting inaccurate information into that input image, as our intention here is to create a strong reference point for the generated images and avoid view inconsistency issues.

Thus, we propose **Ideal Subset Transformation (IST)**. This transformation is carried out by M, which is a latent diffusion model as illustrated in Fig. 4. As aforementioned, M takes an image and a Gaussian noise sample as inputs and transforms that image into an idealized version that resembles an image captured under perfect lighting, camera position, calibration etc. In Fig. 8, we show some of the results produced by M. Graphically, we are transforming the input image from a point on the teal sphere in Fig. 7a to a point on the magenta ring in Fig. 7b. The magenta ring represents an ideal subset of images that are captured under perfect conditions. Any image in this ideal subset is defined to have the following characteristics or conditions:

1. Camera must be at the same height as the center of the human subject.
2. Weak perspective camera is used to prevent none of the body distortion.
3. Even lighting that is neither too bright nor too dark.

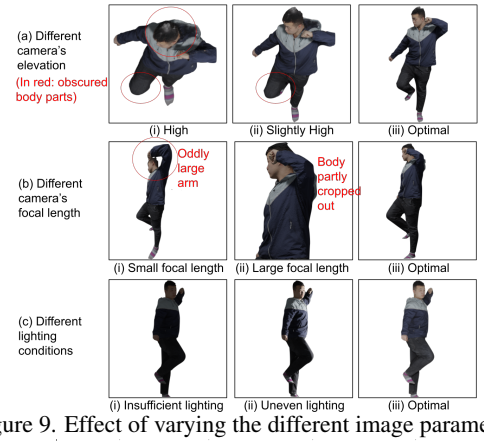


Figure 9. Effect of varying the different image parameters.

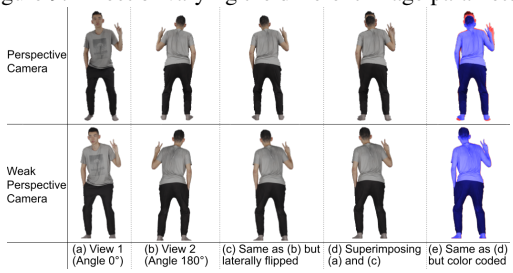


Figure 10. Weak-Perspective camera enables Lateral Inversion

4. All parts of the human subject are within camera frame.

We formally describe M in Eqn. 3:

$$I_0 = M(I_0'), \qquad I_0 \in \mathbb{S} \tag{3}$$

where I_0 is the idealized version of initial input image I_0', and $\mathbb{S}$ is the ideal subset of images.

Our M is trained with paired images (non-ideal images that are captured under random conditions and ideal images captured under the ideal conditions we just defined). The non-ideal image is used as a conditional prior by M, and the corresponding ideal image is the target image that M is asked to recover. Like Zero-1-to-3 [7], our M uses the conditional priors via a combination of cross-attention and channel-concatenation.

3.1.2 Sequential Conditioning (SC)

Sequential Conditioning is the second part of our Cyclic Conditioning Mechanism.

As aforementioned, existing models like Zero-1- to-3 [7] do not take previous image predictions into account when generating subsequent images, resulting in view inconsistency.

In R-Cyclic Diffuser, we resolve this by using **Sequential Conditioning (SC)**. Under SC, our F will first use the ideal image I_0 produced by M to generate I_1. Then, our F will use both I_0 and I_1 to generate I_2. After that, I_0, I_1, and I_2 will be used by F to generate I_3. This is illustrated in Fig. 4. This way, we give our F a means to ensure that $\{I_1, I_2, \dots I_V\}$ will reconcile with one another. In our case, we use $V = 3$. The effect of using and not using SC will be shown in our ablation studies later.

Concretely, SC is formulated as Eqn. 2, and it is implemented by adding I_{t-1}, I_{t-2}, and I_{t-3} as conditional priors to our latent diffusion model F. The conditioning, like Zero-1-to-3 [7], is done via a combination of cross-attention and channel-concatenation.

In every training iteration, a sequence of 4 images $\{I_0, I_1, I_2, I_3\}$ that pertains to the same human subject is randomly selected. The 4 images are all captured in ideal conditions and differ from one another by a 90° azimuth rotation of viewpoint. After training, F will learn to generate an ideal image (I_t) that is a -90° azimuth rotation of I_{t-1}.

To sum up, IST transforms an image from any point within the teal sphere in Fig. 7 to a point on the magenta ring. SC then moves that point to other points on the ring.

3.2. Lateral Inversion Constraint and Cyclic Noise Selection

Naturally, the use of Sequential Conditioning (SC) gives rise to the likelihood of propagation errors. Since if I_1 is incorrectly predicted, possibly due to the randomly sampled Gaussian noise input, then I_2, and later I_3, will likely be incorrectly predicted as well. Thus, in order to address and mitigate the propagation error, we introduce our Lateral Inversion Constraint and Cyclic Noise Selection.

3.2.1 Lateral Inversion Constraint (LIC)

Images that are in our ideal subset $\mathbb{S}$ have to be captured by weak perspective projection. This is an important feature of our ideal subset as it introduces geometric constraints between each pair of images that are 180° (in azimuth) apart. Specifically, if we laterally flip an image out of the pair, then the two images will be pixel-wise aligned (see Fig. 10). We refer to this as a **Lateral Inversion Constraint (LIC)**. With LIC, once an image that is 180° apart from the target image is given as a conditional prior to our F, then F will know which pixels in its predicted target image need to be filled and which ones need to be empty. This greatly reduces the output space of F, thereby simplifying the task given to F. Ultimately, LIC reduces the likelihood and magnitude of prediction error in F, and this in turn mitigates the propagation error for subsequent image predictions.

In terms of implementation, whenever F is given an image that is 180° apart from the target image, we will laterally flip that given image. Formally, instead of $F(I_{t-1}, I_{t-2}, I_{t-3}) = I_t$, we use $F(I_{t-1}, I''_{t-2}, I_{t-3}) = I_t$, where I''_{t-2} is the laterally flipped I_{t-2}. Concretely, to predict I_2, we will use I''_0 instead of I_0. To predict I_3, we use I''_1 instead of I_1. An ablation study on LIC is shown later.

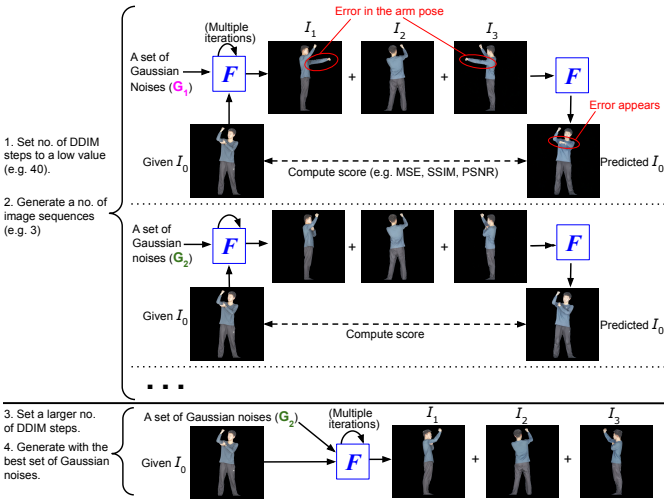


Figure 11. Illustration of our Cyclic Noise Selection.

3.2.2 Cyclic Noise Selection

Besides LIC, Cyclic Noise Selection is another key feature that is designed to mitigate propagation error. As we can see in Fig. 4, each image generated by F is influenced by a Gaussian noise. Since this noise is randomly sampled, this adds unpredictability to the output of F, which sometimes results in prediction and propagation errors. Thus, to ensure that the predictions are always reasonable, we introduce our **Cyclic Noise Selection**, which is used only during testing. Cyclic Noise Selection capitalizes on the concept of cycle consistency and works by first 'completing the cycle' in Fig. 4 i.e. we feed I_1, I_2, and I_3 into F and ask it to generate a predicted I_0. We then compare this predicted I_0 with the initial I_0 given by M using metrics such as MSE, PSNR, and SSIM. The score of the metrics is an indication of how reasonable the sequence of predictions $\{I_1, I_2, I_3\}$ is.

To use Cyclic Noise Selection in our set-up, we will first generate multiple sequences (e.g. 3) of predictions as shown in Fig. 11. While generating these images, it is important to reduce the number of DDIM sampling steps [18] to a low value (e.g. 40) in order to greatly reduce the time required for F to generate each image. In our set-up, reducing DDIM steps from 200 to 40 reduces prediction time by roughly 5 times. Each sequence of images is predicted using a different set of Gaussian noises. Thus, our objective here is to identify which set of Gaussian noises will give us the most reasonable sequence of predictions $\{I_1, I_2, I_3\}$.

To do so, for each sequence, we will compute the error between the I_0 given by M and the I_0 predicted by F using the aforementioned score metrics. We will then pick the set of Gaussian noises (e.g. G_2) that gives the best score. We then re-generate I_1, I_2, and I_3 using G_2 and with the original number of DDIM sampling steps (e.g. 200). According to the authors of DDIM [18], when the number of DDIM steps is reduced, the high-level characteristics of the resulting image will remain unchanged when the same Gaussian noises are used. This is also verified by our Fig. 11.

Using Cyclic Noise Selection, we limit the randomness of our predictions and restrict the prediction and propagation errors caused by bad Gaussian noise samples.

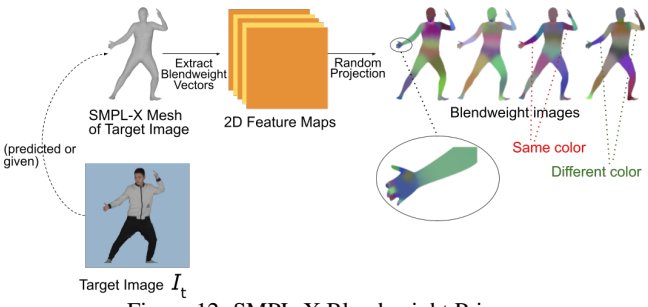


Figure 12. SMPL-X Blendweight Priors

3.3. SMPL-X Blendweight Priors

In 3D clothed human reconstruction, additional priors like a SMPL-X body mesh [10] might be available as an input, but existing models like Zero-1-to-3 and One-2-3-45 did not propose a way to incorporate such priors because they are designed specifically for 3D object reconstruction. To fill this gap, we propose **SMPL-X Blendweight Priors**, which extract human body parts information from a given (or predicted) SMPL-X mesh and then feed this information into F so that F can make more informed predictions.

Every SMPL-X mesh contains a fixed set of vertices V and is associated with the same blendweight matrix B. For each vertex v, matrix B contains a 55-length vector that specifies how v would be displaced when the joints of the SMPL-X body are rotated. This vector, which we refer to as a blendweight vector, can also be interpreted as a vector that determines how much a vertex is associated with every joint in the SMPL-X body. In S-PIFu [1], the authors show that a blendweight vector can be used as a soft human parsing label for a vertex v, and these soft labels outperform discrete ones. Thus, we aim to utilize these blendweight vectors.

We first identify the camera-facing vertices of the given (or predicted) SMPL-X mesh. The camera's position and viewpoint is that of the target image (see Fig. 12). We then retrieve the blendweight vectors of those identified vertices, giving us a set of 2D feature maps (55×H×W) where each pixel location contains a blendweight vector of a vertex.

Then, in order to reduce the number of channels, we apply random projection on the 2D feature maps. Concretely, we apply $S' = RS$ where S are the 2D feature maps that are reshaped to 55×HW, R is a randomly initialized but immutable matrix with shape 3×55, and S' is our blendweight image that has a shape of 3×HW. A blendweight image is interpretable (i.e. can be visualized using RGB channels, as shown in Fig. 12). While a blendweight image reduces the number of channels from 55 to 3, we observe in Fig. 12 that within a blendweight image, there are body parts that are labelled with very similar colors. In order to resolve this,

we generate a total of k blendweight images using k different R. The choice of k is a user-defined hyperparameter that has a trade-off between accuracy and computational cost. In our set-up, we set k=4, giving us 4 blendweight images. Finally, these blendweight images are concatenated together to form our SMPL-X Blendweight Priors. These priors are fed into F as input, as shown in Fig. 5.

By using SMPL-X Blendweight Priors, we provide our F with a precise human parsing map that assigns a different and yet meaningful label to every single location on a human body. This map serves as a template of the target image, showing where the different body parts (e.g. arms, fingers, ears) will be in the target image. The task of F is to predict the target image by filling up this template.

4. Experiments

4.1. Datasets

In our experiments, we use the THuman2.0 dataset [23] as the training set for both our R-Cyclic Diffuser and the existing SOTA models. The THuman2.0 dataset contains 526 body scans (i.e. meshes) of real-life human subjects. We use a 80-20 train-test split. Prior to training, Zero-1-to-3 [7], One-2-3-45 [6], and our R-Cyclic Diffuser are already pre-trained with the same Internet-scale image datasets [17] (>2 billion images) used by Stable Diffusion [13].

For each training mesh, we render RGB images at 10 degree intervals (azimuth rotation), with random elevation angles, lighting conditions, and focal lengths using a perspective camera. We do the same to generate images in the ideal subset $\mathbb{S}$ except that optimal elevation, lighting condition, and a weak-perspective camera is used.

We also use BUFF dataset [24] to evaluate the models. No model is trained using the BUFF dataset. We followed IntegratedPIFu [2] and carried out systematic sampling (using the sequence number) on the dataset, giving us 101 human meshes to be used for our evaluation. Performing systematic sampling allowed us to avoid meshes that have both the same human subject and the same pose.

4.2. Comparison with State-of-the-art

We trained two versions of R-Cyclic Diffuser. The first version assumes that SMPL-X priors are not available and does not use our SMPL-X Blendweight Priors. The second version assumes that SMPL-X priors are available and uses the SMPL-X Blendweight Priors. We compare our two models against existing SOTA models on single-view clothed human reconstruction. The existing models include Zero-1-to-3 [7], One-2-3-45 [6], IntegratedPIFu [2], ICON [20], ECON [21], and D-IF [22]. We use Zero-1-to-3 because it is the pioneer in this new approach to single-view 3D reconstruction. But as shown in Fig. 1, Zero-1-to-3 is designed for object reconstruction and does not work well in human

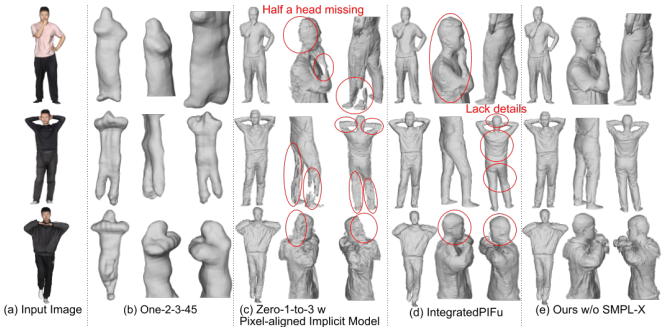


Figure 13. SOTA vs Ours (SMPL-X meshes are not given). See Supp. Mat. for a higher resolution version of this figure.

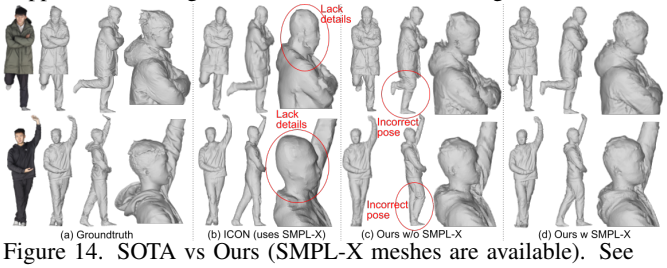


Figure 14. SOTA vs Ours (SMPL-X meshes are available). See Supp. Mat. for a higher resolution version of this figure.

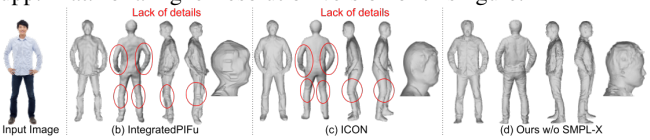


Figure 15. Results on Internet Photos from Shutterstock. See Supp. Mat. for a higher resolution version of this figure.

reconstruction. To improve their performance, we made a slight change to it by using a multi-view pixel-aligned implicit model, rather than their Score Jacobian Chaining [19], to reconstruct 3D meshes from the generated images. This multi-view pixel-aligned implicit model is the same one used in our R-Cyclic Diffuser. Also, any model that uses SMPL-X prior is given the groundtruth SMPL-X prior.

To evaluate 3D clothed human meshes, we followed [2, 4, 15] and used Chamfer Distance (CD) and Point-to-Surface (P2S) as metrics. To evaluate 2D novel views, we followed [7] and used PSNR, SSIM, and LPIPS as metrics.

Qualitative Evaluation We evaluate the models qualitatively in Figs. 2-3 and Figs. 13-14. The figures show that our models, unlike existing methods, can construct meshes with fine details (e.g. clothes wrinkles, nose, ears, hair) in not just front but side and back views as well. Moreover, unlike existing models, our method does not produce artefacts or empty gaps in the reconstructed meshes. We also show our results on real Internet photos in Fig. 15. See our Supp. Mat. for more results, including qualitative comparisons with ECON [21] and D-IF [22].

Quantitative Evaluation Our quantitative results can be seen in Tab. 1, and it shows that our models (both with and without SMPL-X) are able to outperform the existing models in all metrics for both datasets.

Table 1. Quantitative Evaluation with SOTA on 3D Reconstruction

Methods	THuman2.0		BUFF	
	CD (10^{-4})	P2S (10^{-4})	CD (10^{3})	P2S (10^{3})
IntegratedPIFu	5.925	5.777	2.171	2.092
Zero-1-to-3 w Pixel-aligned	6.560	5.093	2.562	2.327
One-2-3-45	8.613	7.556	3.905	4.228
Ours (w/o SMPL-X)	**5.410**	**4.879**	**2.040**	**1.852**
ICON (uses SMPL-X)	0.9934	0.9008	0.8668	0.7912
ECON (uses SMPL-X)	1.108	0.9053	0.9073	0.9774
D-IF (uses SMPL-X)	1.025	0.9149	0.9116	0.7696
Ours (w SMPL-X)	**0.9409**	**0.8806**	**0.7119**	**0.6091**

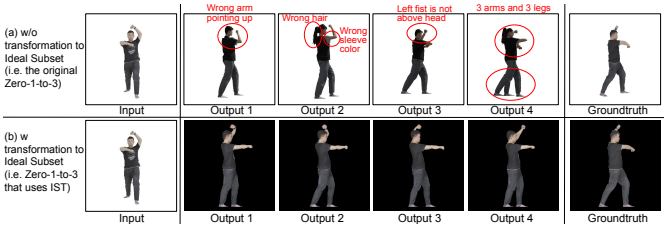


Figure 16. Ablation on our Ideal Subset Transformation (IST).

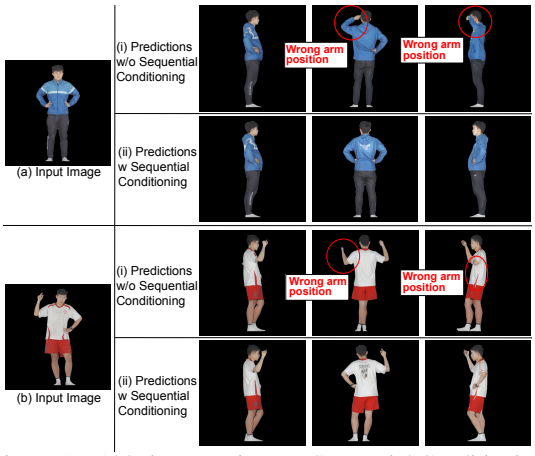


Figure 17. Ablation on using our Sequential Conditioning.

4.3. Ablation Studies

Evaluation of Ideal Subset Transformation (IST) We compare a Zero-1-to-3 that does not use our Ideal Subset Transformation (IST) with a Zero-1-to-3 that does. Using the same input image, we ask both models to produce an output repeatedly and show the results in Fig. 16. We find that without IST, the outputs (views) are clearly less consistent with one another (higher variance) and have a much higher degree of inaccuracy as well.

A quantitative evaluation is shown in Tab. 2, where we compare the predicted images with the groundtruth images. The results show that IST improves prediction accuracy.

Evaluation of Sequential Conditioning (SC) To evaluate the usefulness of using SC, we compare a version of our R-Cyclic Diffuser that does not use SC (i.e. only use the input image as prior) with a version that uses it. The results in Fig. 17 and Tab. 3 show that SC reduces view inconsistency between images and improves accuracy of outputs.

Evaluation of Lateral Inversion Constraint (LIC) We evaluate LIC by comparing a R-Cyclic Diffuser that uses LIC with one that does not. The results in Fig. 18 and Tab. 4 show that LIC improves results by ensuring our predictions are pixel-aligned with image priors that are 180° away.

Table 2. Ablation on Ideal Subset Transformation (IST)

Methods	THuman2.0		
	LPIPS ↓	PSNR ↑	SSIM ↑
w/o IST	0.1134	14.59	0.8539
w IST	**0.0860**	**19.74**	**0.8722**

Table 3. Ablation on Sequential Conditioning (SC)

Methods	THuman2.0		
	LPIPS ↓	PSNR ↑	SSIM ↑
w/o SC	0.0860	19.74	0.8722
w SC	**0.0783**	**20.05**	**0.8736**

Table 4. Ablation on Lateral Inversion Constraint (LIC)

Methods	THuman2.0		
	LPIPS ↓	PSNR ↑	SSIM ↑
w/o LIC	0.0783	20.05	0.8736
w LIC	**0.0632**	**21.24**	**0.8933**

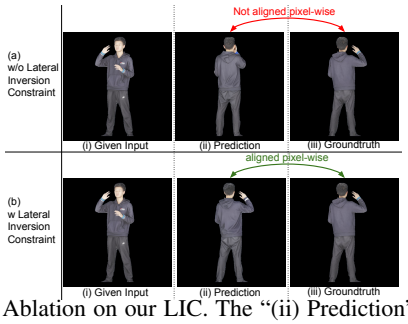


Figure 18. Ablation on our LIC. The "(ii) Prediction" refers to a predicted image that is 180° (in azimuth) from the given input.

Evaluation of SMPL-X Blendweight Priors To evaluate our SMPL-X Blendweight Priors, we compare a R-Cyclic Diffuser that uses SMPL-X Blendweight Priors against one that does not. Qualitatively, we can compare Fig. 14c against Fig. 14d. Quantitatively, we can compare the 4th and 8th rows in Tab. 1. We observe that SMPL-X Blendweight Priors markedly improve results because it aligns the pose of the reconstructed 3D mesh with the groundtruth SMPL-X pose. More results in our Supp. Mat.

5. Conclusion

We have proposed R-Cyclic Diffuser, a framework that adapts Zero-1-to-3's novel approach to clothed human data by fusing it with a multi-view pixel-aligned implicit model. R-Cyclic Diffuser introduced a cyclical conditioning mechanism that is enhanced by two key features - Lateral Inversion Constraint and Cyclic Noise Selection. Moreover, R-Cyclic Diffuser proposed SMPL-X Blendweights Priors, which incorporates a SMPL-X into a latent diffusion model.

Acknowledgements

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Kennard Chan, Guosheng Lin, Haiyu Zhao, and Weisi Lin. S-pifu: Integrating parametric human models with pifu for single-view clothed human reconstruction. In *Advances in Neural Information Processing Systems*, 2022. 6

[2] 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*, pages 328–344. Springer, 2022. 1, 2, 3, 7

[3] Kennard Yanting Chan, Fayao Liu, Guosheng Lin, Chuan Sheng Foo, and Weisi Lin. Fine structure-aware sampling: A new sampling training scheme for pixel-aligned implicit models in single-view human reconstruction. *arXiv preprint arXiv:2402.19197*, 2024. 3

[4] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 535–545, 2021. 7

[5] 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*, pages 300–309, 2023. 2

[6] 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*, 2023. 1, 2, 7

[7] 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*, pages 9298–9309, 2023. 1, 2, 3, 5, 7

[8] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 2

[9] 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. 3

[10] 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)*, pages 10975–10985, 2019. 2, 3, 6

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

[12] 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*, pages 8821–8831. PMLR, 2021. 2

[13] 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*, pages 10684–10695, 2022. 2, 7

[14] 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. 2

[15] 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*, pages 2304–2314, 2019. 2, 3, 7

[16] 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*, pages 84–93, 2020. 2

[17] 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. 7

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[19] 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*, pages 12619–12629, 2023. 1, 7

[20] 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)*, pages 13286–13296. IEEE, 2022. 2, 3, 7

[21] 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*, pages 512–523, 2023. 7

[22] 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*, pages 9122–9132, 2023. 7

[23] 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)*, 2021. 7

[24] 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)*, 2017. 7