FastHuman: Reconstructing High-Quality Clothed Human in Minutes

Lixiang Lin[1] Songyou Peng[2] Qijun Gan[1] Jianke Zhu[1]
[1]Zhejiang University [2]ETH Zurich

Abstract

We propose an approach for optimizing high-quality clothed human body shapes in minutes, using multi-view posed images. While traditional neural rendering methods struggle to disentangle geometry and appearance using only rendering loss, and are computationally intensive, our method uses a mesh-based patch warping technique to ensure multi-view photometric consistency, and sphere harmonics (SH) illumination to refine geometric details efficiently. We employ oriented point clouds' shape representation and SH shading, which significantly reduces optimization and rendering times compared to implicit methods. Our approach has demonstrated promising results on both synthetic and real-world datasets, making it an effective solution for rapidly generating high-quality human body shapes. Project page https://l1346792580123.github.io/nccsfs/

1. Introduction

Human reconstruction is a challenging task due to its high complexity of extreme body poses and sophisticated clothing styles. In general, high-precision laser radar or photometric stereo [62] is required to reconstruct the human body, which greatly increases the cost and only can be done in a controlled environment. Fig. 1 shows our reconstruction results of an in-the-wild video captured by a phone. Our proposed method can reconstruct high-quality human meshes under general lighting environment in a few minutes. By leveraging the power of parametric model [31] and deformation transfer [57] techniques, we are able to generate highly realistic reposed meshes.

With the rapid advancement of neural fields [66], there has been a surge of research devoted to representing 3D geometry and radiance fields using deep neural networks [8, 28, 33, 34, 37, 39, 42, 59, 69, 72, 74]. In these works, 3D geometries are commonly represented using volume density, occupancy, or signed distance functions (SDF). In order to model human avatars, some approaches [6, 43, 45, 60, 61] incorporate the estimated human skeleton and neural rendering to model animatable human avatars in an implicit

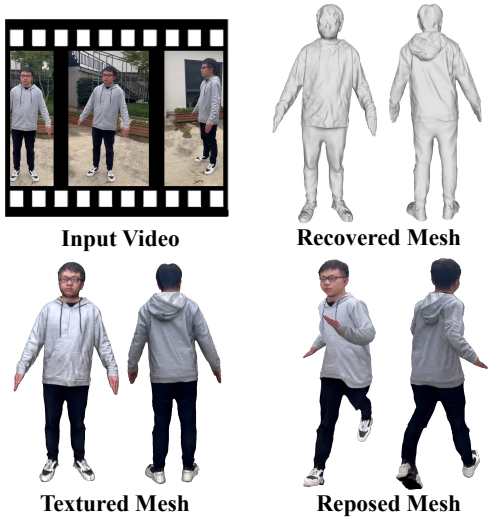


Figure 1. **FastHuman.** Our reconstruction results of an in-the-wild video captured by a phone. Camera parameters are estimated by Colmap [53], and RVM [27] is used for human segmentation. We select an image every 20 frames from the video to construct a set of multi-view images. The whole process can be completed within 10 minutes.

manner. Neural rendering-based methods, while promising, often struggle with reconstructing accurate geometry. This is due to the inherent ambiguity between geometry and appearance, making it challenging to obtain accurate shapes through rendering loss alone. An image can be described by either a plane with complex appearance or a complex geometry with simple appearance. Deep networks can produce smooth surfaces, as the neural network may overfit the color differences between different views. However, shallow networks may lead to local optima due to their poor representation capability. Therefore, it's important to explicitly add multi-view consistency constraints to ensure accurate shape representation.

The implicit Multi Layer Perceptron (MLP) representation is not straightforward, as it requires forward inference to derive geometric information, such as volume density, SDF, etc. It usually takes long time to train these neural rendering-based methods as the whole MLP is updated during each iteration, resulting in slow convergence. Moreover,

the rendering process is computational demanding, since the color of each pixel requires a forward network inference. Although some approaches [10, 35, 47] are proposed to enable real-time rendering, large storage and GPU memory are required as trade off.

To overcome the limitations mentioned above, we introduce a novel coarse-to-fine approach to reconstruct high-quality human meshes from multi-view images. Our approach utilizes an oriented point cloud as shape representation, which allows us to leverage the differentiable Poisson solver [44] for efficient optimization. This ensures that our resulting surfaces are topology-agnostic and watertight, thereby improving the overall quality of the reconstructed mesh. Based on the traditional multi-view stereo approach [11, 53], we warp small patches from the reference frame to source images. We then optimize the shape to ensure local photometric consistency. In addition, we incorporate shape-from-shading (SFS) techniques [73] and estimate the 3rd sphere harmonic (SH) coefficients to represent illumination and jointly refine the shapes and albedos with the shading formulation. As we adopt the simple shading model, the rendering process is sped up substantially compared to the conventional neural rendering methods. In summary, the main contributions of this paper are in the following.

- We present *FastHuman*, a coarse-to-fine pipeline to reconstruct high-quality clothed human bodies from multi-view images in just a few minutes.
- We propose a mesh-based patch-warping strategy to regularize surface optimization in the first stage. In the second stage, we fix the mesh topology and suggest an effective shading-based objective to refine the geometric details further and recover albedos based on shape from shading framework.
- We show the state-of-the-art 3D reconstruction results with significantly reduced computational time compared to existing methods on both synthetic and real-world datasets.
- By taking advantage of parametric model, deformation transfer and SH illumination, we produce realistic reposing and relighting images.

2. Related Works

2.1. Human Reconstruction

Recovering 3D human body shapes from 2D images or videos has been extensively studied for decades [2, 30, 56, 58, 63]. Existing approaches can be roughly divided into two categories: parametric model-based methods and model-free approaches.

Parametric Model-based Human Reconstruction Many research efforts are devoted to building the statistical human body models from 3D scans [3, 18, 31, 40, 67]. In this way, human reconstruction is reduced to the parameter estimation problem, where the shape coefficients and joints transformation are predicted from images [5, 20, 24, 40]. Most of these parametric model-based approaches only produce a naked human body, so the geometries of clothing, hair, and other accessories are typically ignored.

Model-free Human Reconstruction Some approaches reconstruct human body without a predefined model. [50, 51] propose to represent the detailed human by a pixel-aligned implicit function, which predicts the occupancy for any locations. The occupancy for the sampled 3D point can be computed on the fly. With the rapid development of neural rendering, human reconstruction can be also viewed as the by-product of image synthesis. [43, 45, 61] dynamically synthesize the human image. These methods use the SMPL [31] parameters as inputs, and the volume rendering is employed to render images. Coarse human mesh can be extracted from the volume density through Marching cube algorithm [32].

2.2. Multi-View 3D Reconstruction

Traditional Multi-View Stereo (MVS) methods estimate the depth maps by matching feature points across different views. Most of them assume that the appearance of a surface point is consistent in all visible views [11, 53]. Depth fusion and Poisson surface reconstruction [22] are required to extract a watertight mesh, whereas the surface details may be smoothed due to depth fusion. Recently, the learning-based MVS methods have received a lot attentions [16, 68]. DeepMVS [16] follow the traditional pipeline while replace the hand-crafted features to deep features. MVSNet [68] warps deep features into the reference camera frustum to build a cost volume via differentiable homographies.

The recent trend in neural implcit representation and neural rendering also makes an impact in multi-view 3D reconstruction. IDR [69] employs an MLP to represent scenes through SDF and light field implicitly, where color is only calculated at the surface intersection with a ray. [8, 37, 59, 70] incorporates volume rendering and implicit surface representation to learn the geometry from multi-view images. Although they can produce decent object-level 3D reconstruction, they suffer from reconstructing fine geometric details due to the use of a simple color loss, which cannot resolve the shape-appearance ambiguity. Moreover, using an MLP as shape representation leads to very slow optimization speed. In contrast, we propose in this paper two additional constraints for optimization, and use a lightweight point representation to speed up the optimziation.

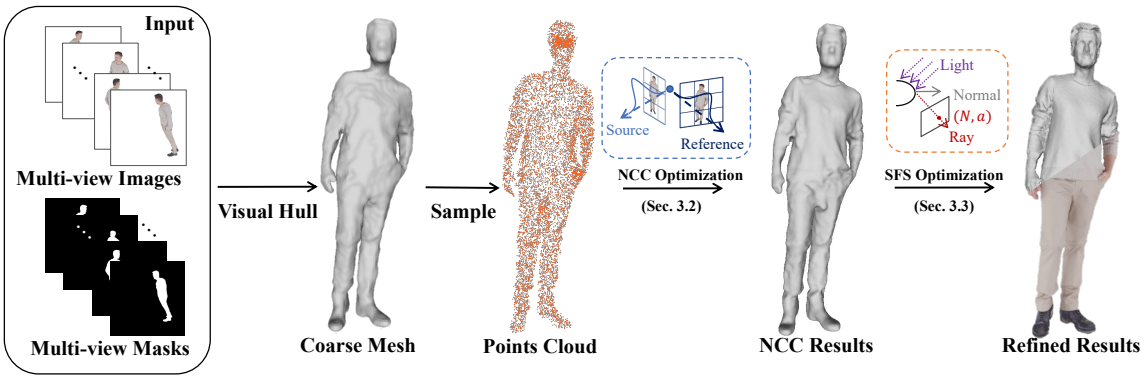


Figure 2. **Method Overview**. The initialized mesh is obtained by visual hull over the provided multi-view human masks. The oriented point clouds are sampled and optimized by multi-view patch-based photometric constraints. Moreover, we fix the mesh topology and employ a shape from shading refinement to refine the coarse mesh and recover the albedo.

2.3. Shape from Shading

Shape-from-shading (SFS) deals with the recovery of shape from a gradual variation of shading in image, which was first proposed by Horn [14]. SFS is an ill-posed problem due to the ambiguity among lighting, reflectance and shape. There are some numerical solutions for SFS such like variational approach [15, 46] and PDE methods [7].

With the prevalence of consumer-level depth cameras, SFS approaches use the rough depth map as initialization for shape refinement [38, 64, 71, 75]. A series of work [12, 13, 41, 52] also incorporate shape from shading and uncalibrated photometric stereo to upsample the low-resolution depth map from a RGB-D sensor in order to match the corresponding RGB image. We take the inspiration from SFS literatures and apply the shading refinement for the multi-view reconstruction task to further disambiguate shape from appearances, so we can recover fine geometric details.

3. Methods

We firstly introduce the oriented point clouds representation in Section 3.1. Next, we introduce our overall pipeline and the patch-based photometric consistency loss in Section 3.2, and the shape refinement from shading in Section 3.3. Details for our implementation are given in Section 3.4.

3.1. Oriented Point Clouds Shape Representation

A recent work [44] introduces a hybrid shape representation called Shape As Points (SAP), where they introduce an efficient differentiable Poisson solver (DPSR) to bridge oriented point clouds, implicit indicator functions, and meshes altogether. Compared to works using neural implicit-based shape representations [29, 36, 37, 59, 69, 72], SAP allows representing any shapes as light-weight oriented point clouds, and yields the high-quality watertight meshes much more efficiently. Therefore, we leverage the power of SAP's optimization-based pipeline as the geometric representation for human reconstruction.

3.2. Multi-view Photometric Consistency

Fig. 2 shows the overview of our proposed coarse-to-fine framework. Given masks of multi-view images, we firstly estimate an initial mesh via visual hull [26]. Next, we sample an oriented point cloud $S = \{\boldsymbol{x} \in \mathbb{R}^3, \boldsymbol{n} \in \mathbb{R}^3\}$ from the inital mesh as the shape representation. During optimization, we generate a watertight mesh via DPSR and differentiable marching cubes (DMC):

$$\chi = \text{DPSR}(S) \tag{1}$$

$$\mathcal{M}(V, F) = \text{DMC}(\tanh(\chi)). \tag{2}$$

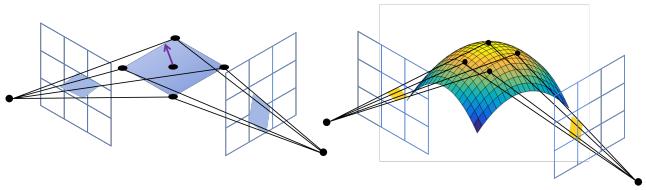
χ represents an indicator function, where 1 indicates inside the object and 0 outside. V and F denote the vertices and faces of the mesh $\mathcal{M}$, respectively. The forward inference of DMC is the generic marching cube algorithm, and the gradients can be effectively approximated by the inverse surface normal [48]. The whole process is fully differentiable, so the loss can be backpropagated to update the oriented point clouds S.

Given the input mesh $\mathcal{M}(V, F)$ with vertices V and faces F, a differentiable renderer [25] denoted as ζ renders the attributes on vertices to pixels given the camera parameter π, which contains intrinsic matrix K and extrinsic matrix T. The rendered silhouette $\hat{M}$ can be obtained by interpolating the constant value of $\mathbf{1}$

$$\hat{M} = \zeta(V, F, \mathbf{1}; \pi). \tag{3}$$

We impose the silhouette loss to limit the boundary of the generated mesh within the mask annotations,

$$\mathcal{L}_{sil} = \sum_{i=1}^{N} ||M_i - \hat{M}_i||_2^2, \tag{4}$$



(a) Point-based patch warping (b) Mesh-based patch warping

Figure 3. **Point-based v.s. Mesh-based Patch Warping**. For point-based patch warping (left), the patch is created by the position of the center pixel and the corresponding normal, which can not represent real geometric information. While for mesh-based patch warping (right), each patch pixel maps to the accurate 3D position through rasterization.

where $||\cdot||_2^2$ represents L_2 norm. $i = 1, \cdots, N$ represents all views.

In order to enforce multi-view photometric consistency, we introduce a mesh-based patch warping strategy. Traditional point-based patch warping relies on the center pixel and normal vector to define a plane, which is then transformed into the source image through homography transformation. However, it is important to note that the resulting plane may not accurately capture the real geometry of the object. By utilizing differentiable Poisson reconstruction and mesh representation, we can determine the accurate 3D position of each pixel within the patch and then transform it into the source image via inverse projection. Fig. 3 illustrates the contrast between the two methods.

We define a patch on an image as $\boldsymbol{p}$, and the color information in the patch is supposed to be consistent among different views. We obtain the exact 3D position corresponding to every pixel through the renderer ζ

$$\hat{P} = \zeta(V, F, TV; \pi) \tag{5}$$

where $\hat{P}$ is the rendered position map. Each valid pixel of $\hat{P}$ represents the corresponding 3D location in the camera coordinate. The third dimension of $\hat{P}$ represents the rendered depth map $\hat{D}$. We warp the patch $\boldsymbol{p}$ on the reference frame to the source frame

$$H_{r\to s}(\boldsymbol{p}) = \pi_s(\pi_r^{-1}(\boldsymbol{p})) \tag{6}$$

$$\hat{P}_s(\boldsymbol{p}) = \mathcal{I}(\hat{P}_s, H_{r\to s}(\boldsymbol{p})), \tag{7}$$

where $\hat{P}_r(\boldsymbol{p})$ and $\hat{P}_s(\boldsymbol{p})$ indicate the 3D position of reference patch and source patch, respectively. we use s and r as subscript to represent source and reference images. $H_{r\to s}(\boldsymbol{p})$ represents the source patch reprojected from reference patch. $\mathcal{I}$ is the bilinear interpolation operation. Grayscale and depth of source patch also can be obtained by the formula similar to Eq. 7.

$$G_s(\boldsymbol{p}) = \mathcal{I}(G_s, H_{r\to s}(\boldsymbol{p})) \tag{8}$$

$$D_s(\boldsymbol{p}) = \mathcal{I}(D_s, H_{r\to s}(\boldsymbol{p})) \tag{9}$$

We convert color images $\{I_i\}$ into grayscale images $\{G_i\}$, and maximize the normalized cross-correlation (NCC) to ensure multi-view photometric consistency

$$\text{NCC}(G_r(\boldsymbol{p}), G_s(\boldsymbol{p})) = \frac{\text{Cov}(G_r(\boldsymbol{p}), G_s(\boldsymbol{p}))}{\sqrt{\text{Var}(G_r(\boldsymbol{p})\text{Var}(G_s(\boldsymbol{p}))}} \tag{10}$$

where Cov is covariance and Var is variance. $G_r(\boldsymbol{p})$ and $G_s(\boldsymbol{p})$ represent the gray value of reference patch and source patch, respectively. NCC scores are computed between the sampled reference and source patches on all source images.

To avoid corresponding patches from occlusion, we compare the the rendered patch depth and the reprojected patch depth, and discard the patches that differ widely. To further guarantee the patches are visible on all source views, we only consider those patches whose NCC scores are above a certain threshold. We impose the multi-view photometric consistency loss on the mesh vertices, and backprop to update the oriented point cloud

$$\mathcal{L}_{\text{ncc}} = \begin{cases} 1 - \text{NCC}(G_r(\boldsymbol{p}), G_s(\boldsymbol{p})) & \delta > 0 \\ 0 \quad \text{else} \end{cases} \tag{11}$$

$$\begin{aligned} \delta = &(|T_s T_r^{-1} K_r^{-1} \boldsymbol{p} - \hat{D}_s(\boldsymbol{p})| < \delta_d) \\ &\times (\text{NCC}(G_r(\boldsymbol{p}), G_s(\boldsymbol{p})) > \delta_{\text{ncc}}) \end{aligned} \tag{12}$$

where δ_d is depth threshold and δ_{ncc} is NCC threshold. $\hat{D}_s(\boldsymbol{p})$ and $T_s T_r^{-1} K_r^{-1} \boldsymbol{p}$ represent the interpolated patch depth and the reprojected patch depth, respectively. K and T represent intrinsic matrix and extrinsic matrix.

3.3. Shape from Shading Refinement

Compared to the initial mesh from visual hull, Multi-view photometric consistency helps to obtain decent geometric details through patch warping. In order to further obtain more fine details on a per-pixel level and recover albedos, we employ SFS refinement discussed as following. In general, the color of human skin and clothes mainly have diffuse reflection, which fits the assumption of shape from shading (SFS) algorithm. Once the mesh is obtained from previous subsection, an SFS refinement is employed to improve the mesh and extract albedos from multi-view images. We firstly review the image formation model using the SH illumination, and then propose the geometry refinement and albedo extraction method in detail.

Image Formation Model When objects in a scene are non-emitters and the light source are infinitely distant, the image irradiance equation can be defined as in [19],

$$B(x, w_o) = \int_\Omega L(w_i)\rho(w_i, w_o)\max(w_i \cdot \boldsymbol{n}_x, 0)dw_i, \tag{13}$$

where $B(x, w_o)$ is the reflected radiance. $x, \boldsymbol{n}, w_i, w_o$ are the spatial location, surface normal, incident light direction

and viewing direction, respectively. The domain of integral Ω is a semi-sphere centered at x. $L(w_i)$ is the light intensity from direction w_i. $\rho(w_i, w_o)$ is the bidirectional reflectance distribution function (BRDF) of the surface. With the assumption of Lambertian Surface, the reflection is constant from all directions of views.

We make use of 3 rd sphere harmonic (SH) coefficients to represent the general lighting. Due to the orthogonality of the SH basis, Eq. 13 can be derived as below

$$B(x) = \rho_x \sum_{i=1}^{3^2} l_i Y_i(\boldsymbol{n}_x) \tag{14}$$

where ρ_x is the albedo at point x. l_i is SH coefficients. Y_i is the SH function determined by the surface normal $\boldsymbol{n}$.

We estimate the l_i according to the mesh from subsection by minimizing the difference between the image density and the computed image irradiances

$$\hat{\boldsymbol{l}} = \arg\min_{\boldsymbol{l}} \sum_{x} || \sum_{i=1}^{n^2} l_i Y_i(\hat{\boldsymbol{N}}(\pi(x))) - G(\pi(x))||_2^2 \tag{15}$$

$$\hat{\boldsymbol{N}} = \zeta(V, F, V_n; \pi) \tag{16}$$

We interpolate the vertex normal V_n to build the normal map $\hat{\boldsymbol{N}}$. For each valid pixel and its corresponding point x, we try to minimize the L_2 norm between the grayscale value and the computed irradiance to obtain the SH coefficients. Since this is an overdetermined problem, we use least squares to estimate SH coefficients.

Once the SH coefficients are estimated, we fix them to refine the coarse mesh and extract albedo. We first extract albedo from captured images, and then refine the albedo and geometry jointly.

$$\hat{A} = \zeta(V, F, V_{\text{albedo}}; \pi) \tag{17}$$

$$\mathcal{L}_{\text{sfs}} = \sum_{x} |\hat{A}(\pi(x)) \sum_{i=1}^{n^2} l_i Y_i(\hat{\boldsymbol{N}}(\pi(x)) - I(\pi(x))| \tag{18}$$

where $\hat{A}$ is the interpolated albedo map. $|\cdot|$ denotes L_1 norm. In order to prevent overfitting or getting stuck at the local optima, we introduce the regularization terms to penalize the surface deformations and texture consistency

$$\begin{aligned} \mathcal{L}_{\text{reg}} &= \mathcal{L}_{\text{mesh}} + \mathcal{L}_{\text{albedo}} \\ &= |LV| + |LV_{\text{albedo}}| \end{aligned} \tag{19}$$

where L denotes the Laplacian matrix. $\mathcal{L}_{\text{reg}}$ forces that the adjacent vertices have similar positions and colors.

3.4. Implementation Details

We firstly extract initial mesh via visual hull at a grid resolution of 128^3. Secondly, we uniformly sample 50k oriented points as our shape representation, and the silhouette and NCC loss in Eq. 4, 11 are backpropagated to update the point cloud. For the patch warping loss $\mathcal{L}_{ncc}$, we preselect 4 adjacent images for each reference image as source images. We perform the optimization at the resolution of 512^3 and the degree of gaussian smoothing $sig = 4$ for 10 epochs. The NCC threshold is $\delta_{\text{ncc}} = 0.5$, and the depth threshold is $\delta_d = 0.01$. The patch size is 11×11. For ablation studies in the number of oriented points and patch size, please refer to supplementary materials. As done in [44], we resample points and normals every other epoch in order to increase the robustness of the optimization process. The weights for the loss terms are $\lambda_{sil} = 20$, $\lambda_{\text{ncc}} = 5$. we use Adam optimizer [23] for optimization, and the learning rate for updating oriented point clouds is $1e^{-3}$. In SFS refinement stage, mesh is exported from point cloud and the topology is fixed. we firstly optimize vertex albedo for 200 epochs at the learning rate of $1e^{-2}$. Then, we refine the mesh and albedo simultaneously for another 100 epochs. The learning rate for vertices and albedo are $1e^{-3}$ and $5e^{-3}$, respectively. The weights for the loss terms are $\lambda_{\text{sfs}} = 20$, $\lambda_{\text{mesh}} = 50$ and $\lambda_{\text{albedo}} = 1$, respectively.

4. Experiments

In this section, we first present the experimental results for reconstructing human body from multi-view images. We compare our proposed method with the current state-of-the-art techniques, and demonstrate that it is also applicable to multi-view videos. We also conduct ablation studies to evaluate the effectiveness of multi-view photometric consistency and shading refinement in our approach.

4.1. Results on Multi-view Images

Since the existing multi-view human datasets do not have ground truth 3D meshes, we render meshes to generate multi-view images for the quantitative comparison. We collect 40 high-resolution photogrammetry scans from RenderPeople [49], and render these meshes using the off-the-shelf software Blender [4]. For each scan, we render 19 images in a circle around the mesh with the resolution of 1024×1024.

We compare our proposed method against recent multi-view human reconstruction [55], multi-view scene reconstruction [8, 34, 59, 69] and colmap [53]. We also compare to NeuS NGP, where we run NeuS with multi-res feature grids from this repo[1] with our data. Since PIFu [50] and PIFuHD [51] are proposed for human reconstruction from a single image, they do not consider the camera information. We do not compare with PIFu and PIFuHD. Similar to PIFu,

[1] https://github.com/bennyguo/instant-nsr-pl

Methods	Normal C.	Chamfer-L_1	PSNR	Optimization Time	Rendering Time
NeRF [34]	0.43	2.32	26.43	126 min	3.0 s
Colmap [53]	0.18	0.36	-	5 min	-
IDR [69]	0.12	0.34	24.46	63 min	2.5 s
NeuS [59]	0.11	0.34	**30.64**	338 min	56 s
NeuralWarp [8]	0.14	0.40	23.39	245 min	63 s
NeuS NGP [35, 59]	0.15	0.48	29.31	8 min	0.05s
DiffuStereo [55]	0.17	0.41	-	**1 min**	-
FastHuman (Ours)	**0.06**	**0.18**	30.58	6 min	**0.01 s**

Table 1. **Quantitative Comparison on Our Synthetic Dataset.** We report a quantitative comparison of 3D reconstruction from multi-view images, and the numbers are average from 40 scans. Compared to baselines, our method can attain better reconstruction quality and also high-quality rendering. Moreover, our optimization speed is significantly faster than neural implicit-based approaches and on part with method with pretraining [55] or optimized traditional pipeline [53]. Moreover, We show the possibility of real-time rendering at 100 FPS.

we adopt three reconstruction performance metrics including normal projection error, Chamfer distance and PSNR.

Table 1 shows the quantitative results. Our proposed approach achieves the lowest Chamfer distance and normal projection error. The PSNR results show that our extracted albedo is able to render the photo-realistic images. We employ Marching Cube on the volume density estimated by NeRF to extract mesh and remove the extra surfaces according to masks. The results of Colmap and Diffustereo are in point clouds form, we employ Screened Poisson surface Reconstruction [21] and remove outliers by masks. Since DiffuStereo has to use DoubleField [54] to generate coarse mesh and normals as input while the implementation of DoubleField is not publicly available. We use the coarse mesh optimized by our proposed multi-view photometric constraints as the input, and fuse the point clouds extracted from all views.

The reconstructed mesh of NeRF is very rough and inaccurate. NeuS renders high-fidelity images. However, it is difficult for rendering loss to handle the ambiguity between appearance and geometry. The reconstructed mesh is different from the ground truth. We use the pre-trained Diffustereo, whose number of cameras and camera positions are different from the training data. Thus, the results of Diffustereo are not satisfied. The reason for poor NeuS NGP performance is the sparsity of input views since feature-grid based methods tend to struggle with this level of ambiguity more than MLP-based methods. Fig. 4 shows the reconstruction results of various approaches.

Our proposed method also has great advantages in terms of computational time on both optimization and rendering. The optimization time for IDR, NeRF, NeuralWarp, NeuS, NeuS NGP are 1 hour, 2 hours, 4 hours 5.5 hours, and 8 min respectively. Since these neural rendering methods use deep neural network as implicit shape representation, it takes lots of time to reach convergence. Although Diffustereo only takes 1 minute for reconstruction after training, DoubleField is needed to generate the coarse mesh as input. Note that DoubleField requires to be pretrained on a large-scale dataset and fine-tuned for 20 minutes. Colmap needs 5 minutes for dense reconstruction, whose reconstruction results are coarse. It takes 5 minutes for our proposed NCC optimization and 1 minute for SFS refinement. Our proposed method has far less computational time on optimization time, which does not require pre-training on large-scale datasets and fine-tuning.

In terms of rendering time, Our approach enables real-time rendering by taking advantage of SH illumination and albedo shading model. Moreover, our reconstructed mesh and texture are compatible with the existing rendering engines such like Blender, Unreal Engine [9], and so on. For the implicit representation-based methods, it takes seconds to render an image, since a forward network inference is required for each valid pixel. Although NeuS NGP can also achieve real-time rendering, it requires much more computations compared to SH shading model. All the experiments about computational time are conducted on the same machine with a single NVIDIA 3090Ti GPU.

4.2. Results on Multi-view Videos

We also conduct experiments on real world captured NHR dataset [65]. NHR dataset is collected by a multi-camera dome system with up to 80 cameras arranged on a cylinder. All cameras are synchronized and capture at 25 frames per second. We conduct experiments on three sequences. We use 24 images with a resolution of 1224×1024 as input. Since there is no ground truth mesh for each frame, we only evaluate the quality of the image synthesis with metric PSNR. We qualitatively compare reconstruction results with IDR [69], NeuS [59] and the point cloud reconstructed using Metashape [1] provided by the dataset. The evaluation metrics are the same as in the previous subsection. for the quantitative and qualitative evaluation of general real-world objects on DTU [17], please refer to the supp. mat.

Figure 4. **Qualitative Comparison on Our Synthetic Dataset.** We show a qualitative comparison of recosntructed surfaces from multi-view images. Compared to baselines, our reconstructions capture most geometric details.

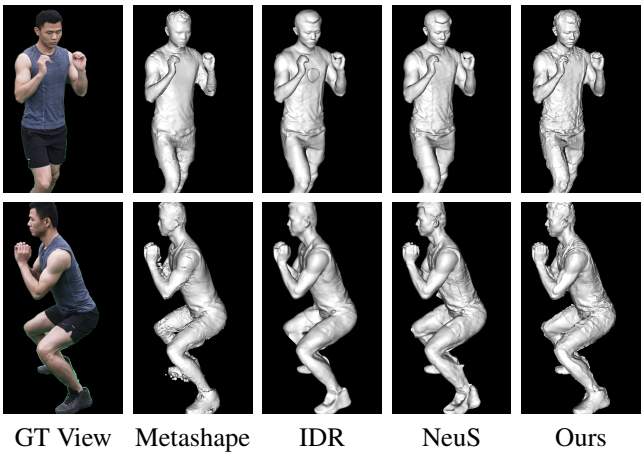


Figure 5. **Qualitative Comparison for Human Reconstruciton on NHR dataset [65].** We compare our method with Metashape [1], IDR [69], and NeuS [59].

Table 2 shows the image synthesis results. It can be seen that we can obtain the photo-realistic images using simple SH illumination and albedos, which greatly accelerate rendering. Fig. 5 shows the reconstruction results. Our proposed method can produce detailed mesh. Similar to the previous subsection, IDR and NeuS tend to produce smooth meshes, hand detail can not be recovered well. Moreover, the ambiguity of geometry and appearance can not be handled by rendering loss alone, which results in wrong reconstruction results of shoes.

	Sport 1	Sport 2	Sport 3
IDR [69]	22.58	21.61	21.68
NeuS [43]	25.52	24.58	24.22
Ours	**25.59**	**24.64**	**24.60**

Table 2. **Image Synthesis Evaluation on NHR Dataset [65] in PSNR.** We conduct experiments on 3 sequences and our method show superior performance compared to other dynamic approach.

We can create a realistic human avatar by combining our reconstructed mesh with SMPL human model [31] and deformation transfer [57]. We pre-select control points on the SMPL mesh and identify the closest points on our reconstructed mesh. By manipulating pose parameters, we can animate the registered SMPL mesh. Deformation can be efficiently transferred to our mesh due to the corresponding transformation of the control points. Additionally, we can synthesize relighting images by replacing the estimated SH coefficients. The results are shown in Fig. 6, demonstrating our ability to generate realistic images with arbitrary lighting and poses.

4.3. Ablation Studies

In this section, we conduct ablation studies on the effect of the proposed NCC loss and SFS loss. We adopt visual

Figure 6. **Relighting and Reposing Results on Our Reconstructed Meshes.** We show qualitative results of relighting and reposing. We register the SMPLX model on the reconstructed meshes. Hand and face parameters are estimated by [40].

Figure 7. **Qualitative Ablation Study on the Proposed Losses**. We show the results of visual hull initialization, only photometric consistency results, only shape from shading results and overall results, respectively.

hull initialization instead of deforming from the sphere for fast convergence. Table 4 shows the quantitative results. As can be noticed, our proposed NCC loss performs better than typical one, since our proposed mesh-based patch warping can obtain accurate 3D position of each pixel. Also, the geometry cannot be recovered correctly with only SFS loss, as visual hull initialization is far from the ground truth. Once the coarse geometry is optimized by NCC loss, SFS loss can finetune the results to get more accurate results. The qualitative results are shown in Fig. 7. Using only NCC loss, we can obtain the coarse results while some details like tie, hands are not recovered well. SFS loss alone may stuck at local optima. Combining NCC loss and SFS loss, we can get accurate results without losing details.

Visual hull	Typical NCC	Our NCC	SFS	Normal	Chamfer
✓				0.15	0.55
✓	✓			0.13	0.32
✓		✓		0.08	0.21
✓			✓	0.14	0.42
✓		✓	✓	**0.06**	**0.18**

Table 3. **Quantitative Ablation Study on the Proposed Loss**. We evaluate the effect of the photometric consistency and shape from shading refinement.

5. Limitations and Conclusions

Our proposed patch warping module, like other MVS methods, may face limitations in accurately reconstructing texture-less objects or under complex lighting conditions. To address this issue, we plan to incorporate deep features for multi-view matching in the future to enhance the module's robustness. Moreover, it's important to note that our SFS module assumes Lambertian surfaces, which may not be suitable for reflective regions.

In this paper, we introduce FastHuman, an efficient coarse-to-fine approach to reconstruct high-quality human bodies from multi-view images. Our method utilizes a mesh-based patch warping optimization technique that leverages orientation point clouds shape representation and multi-view photometric consistency constraints. As the human body and clothing mostly have diffuse characteristics, we employ an image formation model with SH illumination and propose a shading refinement algorithm to recover surface albedo and fine-scale surface detail. To evaluate our approach, we conducted experiments on both synthetic and real-world datasets. Our promising results demonstrated that our approach can reconstruct high-quality human meshes in just a few minutes.

References

[1] Agisoft. Metashape software. *retrieved 20.05*, 2019. 6, 7

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005. 2

[4] Blender, 2018. https://www.blender.org. 5

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 561–578, 2016. 2

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1

[7] Emiliano Cristiani and Maurizio Falcone. Fast semi-lagrangian schemes for the eikonal equation and applications. *SIAM J. Numer. Anal.*, 45(5):1979–2011, 2007. 3

[8] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6250–6259, 2022. 1, 2, 5, 6, 12, 13

[9] Unreal Engine, 2019. https://www.unrealengine.com/. 6

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[11] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2010. 2

[12] Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, and Daniel Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[13] Bjoern Haefner, Songyou Peng, Alok Verma, Yvain Quéau, and Daniel Cremers. Photometric depth super-resolution. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 42(10):2453–2464, 2019. 3

[14] Berthold K. P. Horn. *Shape from shading; a method for obtaining the shape of a smooth opaque object from one view*. PhD thesis, Massachusetts Institute of Technology, USA, 1970. 3

[15] Berthold K. P. Horn and Michael J. Brooks. The variational approach to shape from shading. *Comput. Vis. Graph. Image Process.*, 33(2):174–208, 1986. 3

[16] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2821–2830, 2018. 2

[17] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 406–413, 2014. 6, 12, 13

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2

[19] James T. Kajiya. The rendering equation. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques, (SIGGRAPH)*, pages 143–150, 1986. 4

[20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 2

[21] Michael M. Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29:1–29:13, 2013. 6

[22] Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing (SGP)*, pages 61–70, 2006. 2

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*, 2015. 5

[24] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2

[25] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Trans. Graph.*, 39(6), 2020. 3

[26] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 16(2):150–162, 1994. 3

[27] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proc. of Winter Conf. on Applications of Computer Vision (WACV)*, pages 3132–3141, 2022. 1

[28] Stefan Lionar, Daniil Emtsev, Dusan Svilarkovic, and Songyou Peng. Dynamic plane convolutional occupancy networks. In *Proc. of Winter Conf. on Applications of Computer Vision (WACV)*, 2021. 1

[29] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[30] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Trans. Vis. Comput. Graph.*, 2009. 2

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1, 2, 7

[32] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques, (SIGGRAPH)*, pages 163–169, 1987. 2

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 1, 5, 6, 7

[35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 6

[36] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[37] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1, 2, 3

[38] Roy Or-El, Guy Rosman, Aaron Wetzler, Ron Kimmel, and Alfred M. Bruckstein. Rgbd-fusion: Real-time high precision depth recovery. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5407–5416, 2015. 3

[39] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[40] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 8

[41] Songyou Peng, Bjoern Haefner, Yvain Quéau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proc. of the IEEE International Conf. on Computer Vision Workshops (ICCVW)*, 2017. 3

[42] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 1

[43] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 14294–14303, 2021. 1, 2, 7

[44] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3, 5

[45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[46] Yvain Quéau, Jean Mélou, Fabien Castan, Daniel Cremers, and Jean-Denis Durou. A variational approach to shape-from-shading under natural illumination. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 342–357, 2017. 3

[47] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[48] Edoardo Remelli, Artem Lukoianov, Stephan R. Richter, Benoît Guillard, Timur M. Bagautdinov, Pierre Baqué, and Pascal Fua. Meshsdf: Differentiable iso-surface extraction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[49] Renderpeople, 2018. https://renderpeople.com/3d-people. 5

[50] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2, 5

[51] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 2, 5

[52] Lu Sang, Bjoern Haefner, and Daniel Cremers. Inferring super-resolution depth from a moving light-source enhanced rgb-d sensor: a variational approach. In *Proc. of Winter Conf. on Applications of Computer Vision (WACV)*, 2020. 3

[53] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 501–518, 2016. 1, 2, 5, 6, 12

[54] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 15851–15861, 2022. 6

[55] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In

Proc. of the European Conf. on Computer Vision (ECCV), 2022. 5, 6, 7

[56] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE computer graphics and applications*, 2007. 2

[57] Robert W. Sumner and Jovan Popovic. Deformation transfer for triangle meshes. *ACM Trans. Graph.*, 23(3):399–405, 2004. 1, 7

[58] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM SIGGRAPH Asia*, 2009. 2

[59] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 27171–27183, 2021. 1, 2, 3, 5, 6, 7, 12, 13

[60] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 1

[61] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 1, 2

[62] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 1

[63] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2011. 2

[64] Chenglei Wu, Michael Zollhöfer, Matthias Nießner, Marc Stamminger, Shahram Izadi, and Christian Theobalt. Real-time shading-based refinement for consumer depth cameras. *ACM Trans. Graph.*, 33, 2014. 3

[65] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691, 2020. 6, 7

[66] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, 2022. 1

[67] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[68] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 785–801, 2018. 2

[69] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3, 5, 6, 7, 12, 13

[70] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[71] Lap-Fai Yu, Sai Kit Yeung, Yu-Wing Tai, and Stephen Lin. Shading-based shape refinement of RGB-D images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1415–1422, 2013. 3

[72] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3

[73] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 1999. 2

[74] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[75] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Trans. Graph.*, 34(4):96:1–96:14, 2015. 3

FastHuman: Reconstructing High-Quality Clothed Human in Minutes

Supplementary Material

Figure 8. Example of our synthesized dataset. Our synthesized dataset can provide realistic images.

	Normal C.	Chamfer-L_1
baseline	0.08	0.21
10K points	0.09	0.21
100K points	0.08	0.21
sig = 1	0.10	0.24
sig = 10	0.13	0.31
5×5 patch size	0.09	0.22
15×15 patch size	0.08	0.22

Table 4. Quantitative Ablation Study on the parameter. We evaluate the effect of the number of oriented points and patch size.

6. Ablation Studies

In this section, we present our synthesized dataset and results of ablation studies. Fig 8 shows the images of our synthesized dataset. It can be seen taht our synthesized dataset provide realistic images. We conduct ablation studies on the number of oriented points, sig and patch size. In our paper we use 50K oriented points, sig=4 and the patch size is 11×11. Table 4 shows that our proposed module is robust to the parameters. Different parameters have little effect on the reconstruction results.

7. Results on DTU Dataset

In this section, we give the reconstruction results of our proposed patch warping loss on general objects. We have conducted experiments on DTU dataset [17], compared the reconstruction results against IDR [69], NeuS [59], and NeuralWarp [8]. We employ colmap [53] to obtain the sparse points as additional supervision. Fig. 9 shows the results of Visual Hull and sparse points generated by Colmap [53]. The results of Visual Hull is not correct due to the views of DTU dataset are only from one side. The sparse points generated by Colmap is too sparse to evaluate. For dense reconstruction results of Colmap on DTU dataset, you can refer to IDR [69] for more information. As the objects in DTU dataset have specular reflections, we do not perform shape from shading refinement. Quantitative results are shown in Table 5. It can be seen that our proposed patch warping loss achieves the best Completeness distance and comparable Chamfer distance against NeuralWarp. Qualitative results are presented in Fig. 10. As mentioned in our paper, it takes 15 minutes for our proposed patch warping to reconstruct an object of DTU dataset while other methods requires several hours to obtain the results.

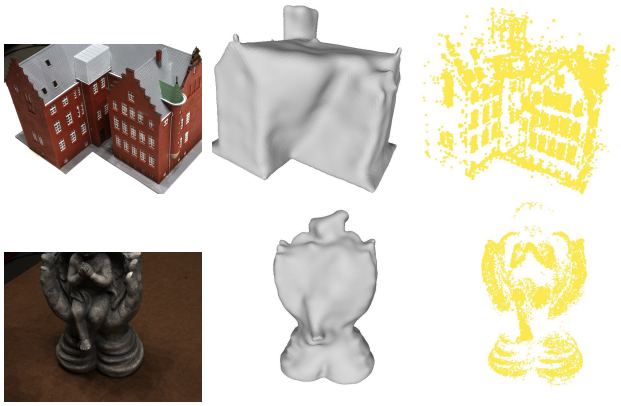
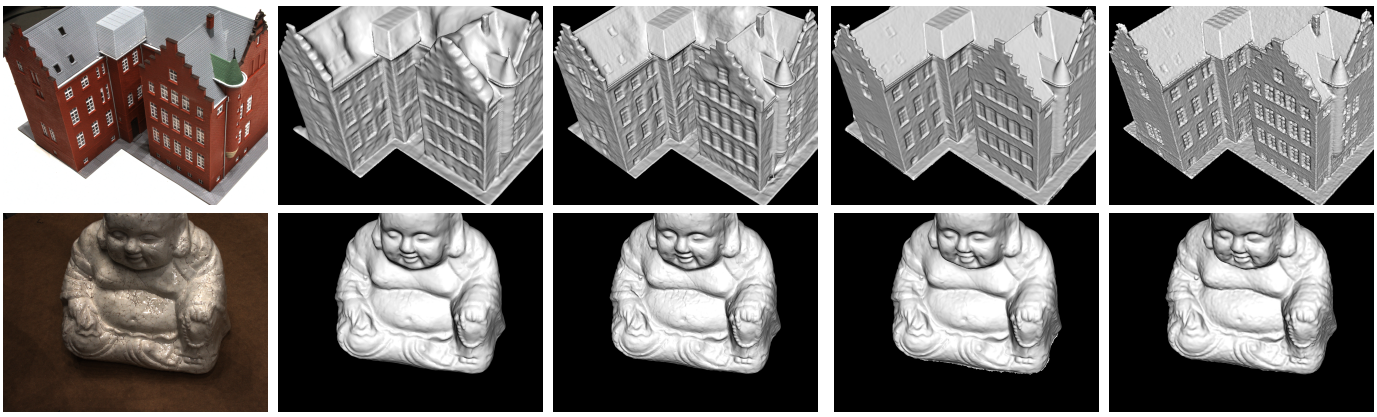


Figure 9. Examples on DTU dataset. We present images, Visual Hull results and sparse points generated by Colmap [53].

Table 5. Quantitative results on DTU dataset. Our method achieves comparable mean Chamfer distance against NeuralWarp [8]

	Accuracy (mm)				Completeness (mm)				Chamfer (mm)			
	IDR [69]	NeuS [59]	NeuralWarp [8]	Ours	IDR [69]	NeuS [59]	NeuralWarp [8]	Ours	IDR [69]	NeuS [59]	NeuralWarp [8]	Ours
24	1.76	0.90	**0.52**	0.81	1.50	0.75	**0.46**	0.48	1.63	0.83	**0.49**	0.64
37	2.16	1.09	**0.82**	1.36	1.55	0.88	**0.61**	0.81	1.86	0.98	**0.71**	1.09
40	0.65	0.58	**0.39**	0.38	0.61	0.54	0.37	**0.31**	0.63	0.56	0.38	**0.35**
55	0.57	0.40	0.40	**0.35**	0.37	0.34	0.37	**0.27**	0.47	**0.37**	0.38	**0.31**
63	1.43	1.62	**1.01**	1.42	0.63	0.64	0.58	**0.43**	1.03	1.13	**0.79**	0.93
65	0.88	**0.68**	0.81	0.75	0.69	**0.51**	0.81	0.76	0.78	**0.59**	0.81	0.75
69	0.88	**0.68**	0.92	0.70	0.66	0.52	0.72	**0.49**	0.77	**0.60**	0.82	**0.60**
83	1.10	1.33	**0.85**	1.11	1.55	1.57	1.55	**0.95**	1.32	1.45	1.20	**1.01**
97	1.31	1.07	**0.84**	1.06	0.99	**0.84**	1.28	1.44	1.15	**0.95**	1.06	1.27
105	0.86	0.78	**0.59**	0.64	**0.66**	0.78	0.77	**0.66**	0.76	0.78	0.68	**0.65**
106	0.71	**0.53**	0.57	0.64	0.60	**0.52**	0.74	0.66	0.66	**0.52**	0.66	0.65
110	1.09	1.71	**0.92**	1.06	0.68	1.16	0.56	**0.41**	0.89	1.44	0.74	**0.73**
114	0.45	0.34	0.41	**0.32**	0.38	0.38	0.40	**0.34**	0.41	0.36	0.41	**0.33**
118	0.55	**0.48**	0.73	0.52	0.46	**0.43**	0.54	0.44	0.50	**0.45**	0.63	0.48
122	0.71	0.57	0.55	**0.54**	0.43	0.41	0.46	**0.34**	0.57	0.49	0.51	**0.44**
Mean	1.01	0.85	**0.69**	0.78	0.78	0.68	0.68	**0.59**	0.90	0.77	**0.68**	**0.68**



GT View | IDR [69] | NeuS [59] | NueralWarp [8] | Ours

Figure 10. **Qualitative Comparison on DTU dataset [17].** We compare our method with IDR [69], NeuS [59], and NeuralWarp [8].