HISR: Hybrid Implicit Surface Representation for Photorealistic 3D Human Reconstruction

Angtian Wang[1*]**, Yuanlu Xu**[2]**, Nikolaos Sarafianos**[2]**, Robert Maier**[2]**, Edmond Boyer**[2]**, Alan Yuille**[1]**, Tony Tung**[2]

[1]Johns Hopkins University
[2]Meta Reality Labs Research

Abstract

Neural reconstruction and rendering strategies have demonstrated state-of-the-art performances due, in part, to their ability to preserve high level shape details. Existing approaches, however, either represent objects as implicit surface functions or neural volumes and still struggle to recover shapes with heterogeneous materials, in particular human skin, hair or clothes. To this aim, we present a new hybrid implicit surface representation to model human shapes. This representation is composed of two surface layers that represent opaque and translucent regions on the clothed human body. We segment different regions automatically using visual cues and learn to reconstruct two signed distance functions (SDFs). We perform surface-based rendering on opaque regions (*e.g.*, body, face, clothes) to preserve high-fidelity surface normals and volume rendering on translucent regions (*e.g.*, hair). Experiments demonstrate that our approach obtains state-of-the-art results on 3D human reconstructions, and also shows competitive performances on other objects.

Introduction

Realistic and accurate reconstruction of geometry and appearance of digital humans has received significant attention over the past few years with applications ranging from creating vivid characters to virtual assistants in customer service, and social telepresence (Lombardi et al. 2018).

Exist techniques recover the geometry and appearance of humans from images (*e.g.*, using multi-view stereo, shape-from-X). Following the introduction of NeRF (Mildenhall et al. 2020), which employs a neural network to capture color and opacity in a 3D volume, methods using neural radiance fields have gained significant popularity over the past few years. The NeRF network learns to estimate the radiance optimally from any viewpoint, resulting in images with photorealism. However, NeRF and the related works struggle to accurately capture fine-level surface details and may generate erroneous discrete floating volumes. Signed distance fields (SDF) (Park et al. 2019a), that model the closest distances to surfaces, have been proposed as an alternative to opacity in implicit shape representations. Their advanced ability

to help recover better geometries has been demonstrated in *e.g.*, (Yariv et al. 2020, 2021; Oechsle, Peng, and Geiger 2021).

Our experiments show that concurrent SDF-based representations lack the ability to model fine structures and high-frequency regions of humans (*e.g.*, hair and complex cloth patterns). While NeRF-like representations are too sensitive to noise and misalignment, and tend to generate large amount of floating volumes on the final reconstruction. Such differences potentially stem from, first, the nature of implicit representation with distance fields which provides a smooth transition from positive to negative values across the surface boundary. SDFs are thus less likely to produce floating volumes, which requires change of the surface direction in high frequency. Second, the Eikonal loss $\|\nabla\Phi(\mathbf{x})\| - 1$, which further enforces the smoothness of object boundaries, and reduces therefore the sharpness and floating volumes.

Previous works (Oechsle, Peng, and Geiger 2021; Yariv et al. 2021; Wang et al. 2021) attempt to combine SDF and NeRF by replacing density fields with SDFs and employing SDF-to-density functions, while performing volume rendering by sampling the entire space. These approaches offer advantages such as enforcing surface smoothness through geometry regularization and enabling training without a segmentation mask. However, we observe that such models lose the crucial capability of NeRF models to capture intricate geometries, like hair strands, and struggle to converge under challenging scenarios. In this paper, we propose a complementary approach that retains the strengths of both SDF and NeRF models by carefully controlling the model's behavior according to different body parts, allowing for improved performances and the ability to capture detailed geometries.

In this paper, we propose a new neural rendering framework for multi-view reconstruction of real humans. We use the signed distance field (SDF) to model shape surfaces and introduce a novel volume rendering scheme to learn a two-layer implicit surface-based representation. Specifically, by introducing a density distribution induced by the SDF, we can perform volume rendering to learn an implicit SDF representation and thus obtain both an accurate surface representation, benefiting from the neural SDF model, and a robust network training in the presence of abrupt depth changes, as enabled by the volume rendering. We performed in-depth quantitative evaluations of our Hybrid Implicit Surface Representation
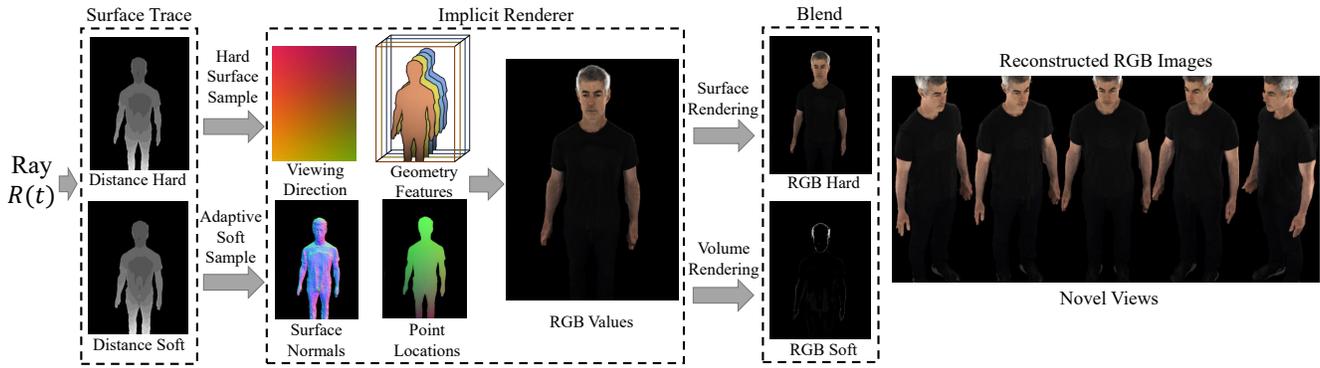
Figure 1: Overview of HISR, which takes viewing rays as input and simultaneously conducts surface and volume rendering. The process is automatically controlled by two SDFs. Once each rendering is done, we blend the colors for the final output. Due to privacy reason, we mask out part of the face.

(HISR) on stage-captured real people and photo-realistic virtual humans, as well as objects from the DTU dataset (Aanæs et al. 2016). We demonstrate that our method is capable of reconstructing photo-realistic 3D clothed humans and clearly outperforms several state-of-the-art approaches quantitatively and qualitatively. Our contributions include:

- A study of state-of-the-art 3D reconstruction approaches for 3D humans, which shows *existing approaches cannot obtain both smooth surface reconstruction and high-fidelity geometry details simultaneously*.

- *A new hybrid representation*, which is surface-based while enabling volume rendering for fine-grained geometry.

- A *computed expectation of SDF values within a conical frustum* when computing the volume densities by considering the viewing rays as cones, which significantly improves the reconstructed geometry.

- *A new loss to regularize the specularity* changes upon viewing directions.

Related Work

Neural Implicit Representations based on ray tracing volume densities (Kajiya and Herzen 1984) formulate the volume rendering process as the solution for the scatter equation under the low albedo assumption. Different from early approaches on volume rendering, which represent objects using explicit primitives (Westover 1990; Zwicker et al. 2001; Wang et al. 2022a), NeRF (Mildenhall et al. 2020) represents objects via implicit functions of volume densities, which has shown high-quality results in the novel view synthesis task. Follow-up works (Barron et al. 2021, 2022) further improve the novel view synthesis ability of NeRF with more fine-grained details. Recent works explore broader applications of NeRF (Chen et al. 2022; Xu et al. 2022; Wang et al. 2022b; Gao, Cao, and Shan 2023; Cai et al. 2023; Wang et al. 2023b). However, the reconstructed geometry of NeRF results in artifacts since the geometry representation lacks surface constraints. Orthogonal to the above studies, works focus on applying implicit representations to human sequences (Muller et al. 2022; Wang et al. 2022c; Işık et al. 2023).

Neural 3D Human Reconstruction has shown great potential in many digital human and AR/VR applications (Mescheder et al. 2019; Park et al. 2019b; Chen and Zhang 2019; Huang et al. 2020). Different from regular objects (Wang, Kortylewski, and Yuille 2021; Wang et al. 2023a), the geometries of human are articulate with large variant appearance on different regions. One of the first approaches to adopt the implicit function representation for 3D human reconstruction from a single image is PIFu (Saito et al. 2019, 2020). PIFu leverages pixel-aligned image features rather than global features. Local details present in the input image are preserved as the occupancy of any 3D point is predicted. Alldieck *et al* (Alldieck, Zanfir, and Sminchisescu 2022) improved upon PIFu by introducing a network that estimates 3D geometry, surface albedo and shading from a single image in a joint manner. In contrast, inspired by the literature of image-based super-resolution, SuRS (Pesavento et al. 2022) demonstrates that fine-scale detail can be recovered even from low-resolution input images using a multi-resolution learning framework. While other approaches using NeRF-type approach for reconstruct 3D human in motion (Pumarola et al. 2021; Gafni et al. 2021; Peng et al. 2021; Noguchi et al. 2021; Jiang et al. 2022; Weng et al. 2022).

Geometry Representation of Signed Distance Function allows for modeling objects by an implicit surface (Park et al. 2019a). IDR (Yariv et al. 2020) proposed to use a single implicit hard surface as object representation and achieved high-quality geometry reconstructions. Follow-up works (Yariv et al. 2021; Wang et al. 2021; Yariv et al. 2023) learn and render SDF functions in a volume rendering manner, which demonstrate comparable better reconstruction for the smooth surface without mask supervision. However, compared to general objects for which getting very accurate segmentation or matting masks might be challenging, getting high-quality human masks is a well-researched topic. Powerful pre-trained segmentation models for clothed humans allow us to obtain the mask at almost no cost. On the other hand, such approaches lose the ability to model geometries with fine-level details, while in this work, we follow the SDF-based approach for geometry reconstruction but with the capacity for modeling detailed geometries.
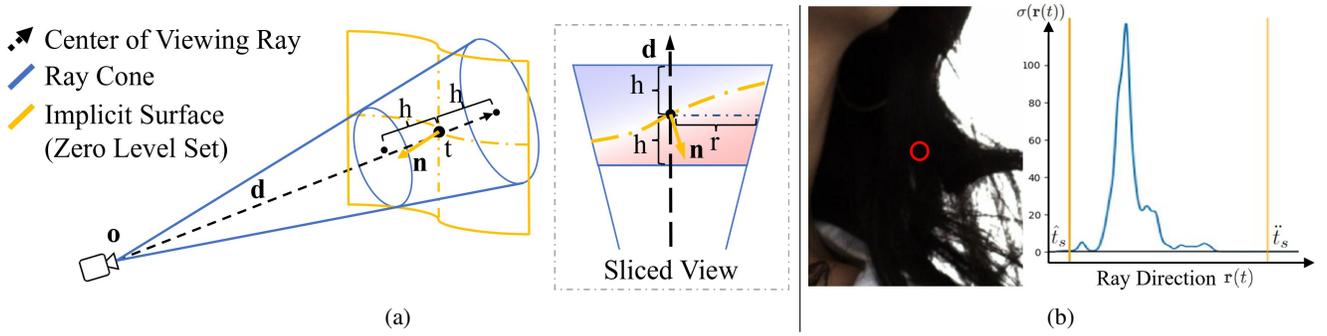
Figure 2: (a) Illustration of cone cast for volume rendering. We assume a uniform SDF inside the conical frustum and compute a close-form solution for the expectation of the SDF values. (b) Examples of volume densities computed from SDF values on hair regions. The ray is sampled at the center of the red circle. We conduct volume rendering from $t = \hat{t}_s$ to $t = \ddot{t}_s$.

Method

Our approach builds on two key components: i) signed distance functions (SDF) (Park et al. 2019a) for representing the instance's geometry, and ii) an implicit neural renderer capable of modeling textures and illumination. To elucidate our method, we first introduce the two primary rendering mechanisms used for implicit representations: surface-based and volume-based. And to better integrate the surface and volume rendering approaches, we introduce a novel technique called integrated SDF for volume densities. This approach enables the seamless fusion and synchronization of updated SDF and volume densities within our approach, enhancing the overall quality and accuracy of the reconstructed human model.

Hybrid Implicit Surface Representation

Surface Rendering. The 3D object is represented as a set of points on surface, which is described as the zero level set of the SDF Φ represented by a neural network (MLP),

$$\mathcal{S} = \left\{ \mathbf{x} \in \mathbb{R}^3 \mid \Phi(\mathbf{x}) = 0 \right\}. \quad (1)$$

During the rendering process, we compute a ray $\mathbf{r}(t) = \mathbf{o} + t \cdot \mathbf{d}$ for a given image pixel, where $\mathbf{o}$ represents the camera location, $\mathbf{d}$ is the viewing direction and t is the depth along the viewing ray. To determine the intersection of the ray $\mathbf{r}(t)$ with the surface $\mathcal{S}$, we use surface tracing to search along the ray for the first zero point, denoted by $\hat{t}$, where $\Phi(\mathbf{r}(\hat{t})) = 0$. The final color observation is computed at the intersection point as:

$$\mathbf{C} = M(\hat{\mathbf{x}}, \hat{\mathbf{n}}, \mathbf{d}, f), \quad (2)$$

where $\hat{\mathbf{x}} = \mathbf{r}(\hat{t})$ is the point location, $\hat{\mathbf{n}}$ is the surface normal at $\hat{\mathbf{x}}$, M denotes the implicit neural renderer implemented with a MLP, and f is a feature vector describing the geometry that is obtained from the SDF network.

Volume Rendering. The 3D object is represented as semi-transparent volumes with volume density $\sigma(\mathbf{x})$ at each location in the scene $\mathbf{x} \in \mathbb{R}^3$. During the rendering process, for each viewing ray $\mathbf{r}(t) = \mathbf{o} + t \cdot \mathbf{d}$, a set of points $\mathbf{r}(t_k)$ is sampled and stored, along with the computed color $\mathbf{c}_k$ and density σ_k. The final color observation is obtained by discretizing the sum, approximating the integral within the volume rendering function (Kajiya and Herzen 1984; Mildenhall et al. 2020):

$$\mathbf{C} = \sum_k T_k \left(1 - \exp\left(-\sigma_k \left(t_{k+1} - t_k \right) \right) \right) \mathbf{c}_k,$$

$$\text{with } T_k = \exp\left(-\sum_{k'<k} \sigma_{k'} \left(t_{k'+1} - t_{k'} \right) \right), \quad (3)$$

where T_k is the transmittance function which encodes the visibility at each sampled point. In order to conduct volume rendering using an SDF as the geometry representation, a signed distance to volume density function Φ is introduced:

$$\sigma(\mathbf{x}) = \Psi(-\Phi(\mathbf{x})), \quad (4)$$

where Ψ is the derivative of the Cumulative Distribution Function (CDF) of Laplace (Yariv et al. 2021) or Gaussian (Wang et al. 2021) distribution.

Hybrid Implicit Surface Representation. In our proposed approach HISR, the geometry is represented as a set of surfaces with volumes filled in between specific surfaces. We assume that the space inside the hard surface $\mathcal{P}_h$ is filled with opaque materials, while the outside volumes are translucent with volume densities $\sigma(\mathbf{x})$. Specifically, the boundary of the translucent region is determined by the hard surface, as the inner boundary, and another surface as the outer boundary namely the soft surface $\mathcal{P}_s$. Whereas the space outside the soft surface is vacant. The inner and outer surfaces are represented by two SDFs, namely hard SDF and soft SDF. The hard s_h and soft SDF values s_s at each location $\mathbf{x}$ in the space $\mathbf{x} \in \mathbb{R}^3$ are:

$$s_h = \Phi_h(\mathbf{x}), \quad s_s = \Phi_s(\mathbf{x}), \quad (5)$$

where Φ_h and Φ_s are two outputs generated by a shared MLP. In HISR, similar to other SDF-based volume rendering approaches, the volume density is represented by the SDF-to-density function $\sigma(\mathbf{x}) = \Psi(\Phi(\mathbf{x}))$.

Hybrid Rendering. Figure 1 illustrates the rendering process of HISR. Given a viewing ray $\mathbf{r}(t) = \mathbf{o} + t \cdot \mathbf{d}$, we perform surface tracing to find the intersection point on the ray with $\mathcal{P}_h$ and $\mathcal{P}_s$. This process involves searching for the zero level along $\mathbf{r}(t)$ to obtain $\hat{t}_s$ and $\hat{t}_h$ such that $\Phi_s(\mathbf{r}(\hat{t}_s)) = 0$ and
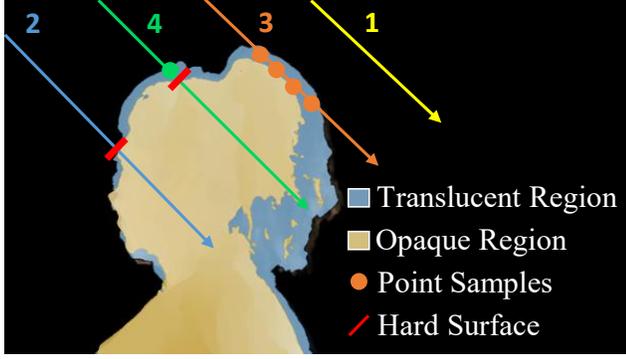
Figure 3: HISR conducts volume rendering in translucent regions and surface rendering on hard surface. The numbers refer to the rendering cases presented in the table below.

$\Phi_h(\mathbf{r}(\hat{t}_h)) = 0$. Consequently, we can deduce that the ray passes through a vacancy from the camera location at $t = 0$ to $t = \hat{t}_s$, which contributes nothing to the final viewing color. From $t = \hat{t}_s$ to $t = \hat{t}_h$, the ray traverses translucent regions, and the final viewing color $\mathbf{C}_s$ is computed using Eq. 3 (the color network is shared with the outer surface rendering). At $t = \hat{t}_h$, the ray intersects with $\mathcal{P}_h$, and the color $\mathbf{C}_h$ can be computed using Eq. 2. Beyond $t = \hat{t}_h$, the ray becomes entirely invisible as the transmittance $T(\mathbf{r}(t))$ reduces to 0. The final viewing color is calculated as $\mathbf{C} = \mathbf{C}_s + T(\hat{t}_h) \cdot \mathbf{C}_h$.

Indeed, the viewing ray $\mathbf{r}(t)$ may only intersect one of the surfaces or even none of them. In the case where $\mathbf{r}(t)$ intersects $\mathcal{P}_s$ but not $\mathcal{P}_h$, we search for a secondary intersection of $\mathbf{r}(t)$ with $\mathcal{P}_s$ such that $\Phi_s(\mathbf{r}(\ddot{t}_s)) = 0$, where $\ddot{t}_s > \hat{t}_s$. This allows the ray to pass inside the volume from $t = \hat{t}_s$ to $t = \ddot{t}_s$ before encountering a vacancy. Figure 3 illustrates all possible cases, which summarize as follows:

$\mathbf{r}(t) \cap \mathcal{P}_s$	$\mathbf{r}(t) \cap \mathcal{P}_h$	start	end	final color
$= \emptyset$	$= \emptyset$	—	—	$\mathbf{C}_{bg}$
$= \emptyset$	$\neq \emptyset$	—	—	$\mathbf{C}_h$
$\neq \emptyset$	$= \emptyset$	$t = \hat{t}_s$	$t = \ddot{t}_s$	$\mathbf{C}_s + T(\ddot{t}_h) \cdot \mathbf{C}_{bg}$
$\neq \emptyset$	$\neq \emptyset$	$t = \hat{t}_s$	$t = \hat{t}_h$	$\mathbf{C}_s + T(\hat{t}_h) \cdot \mathbf{C}_h$

Table 1: Possible cases in ray tracing.

Additionally, we observed that sampling the same number of points on rays belonging to cases 3 and 4 is inefficient. This is because rays in case 4 typically have a much larger sampling range, and the overall rendering quality primarily relies on the largest sampling interval. To mitigate this issue and minimize the maximum sampling interval while staying within the constraints of limited GPU memory, we propose an *Adaptive Sampling Strategy*. This strategy allows for different numbers of points to be sampled on rays while ensuring that the sampling interval on each ray remains similar. We implement this strategy using CUDA as differentiable PyTorch functions. For more details refer to Appendix.

Integrated SDF for Volume Densities

Gaussian Mixture for SDF-to-density Function. The density of semi-transparent volumes is determined using the SDF-to-density functions $\sigma(\mathbf{x}) = \Psi(-\Phi(\mathbf{x}))$. However, we have observed that the Ψ function used in previous works (Yariv et al. 2021; Wang et al. 2021) yields a smooth surface while being successful at capturing intricate geometric details. To address this issue, we introduce the learnable Gaussian mixture model for SDF-to-density:

$$\sigma(\mathbf{x}) = \sum_{k=0}^{K} \alpha_k \cdot \exp(-\frac{(\Phi(\mathbf{x}) - \mu_k)^2}{\beta_k^2}), \qquad (6)$$

where α_k, β_k and μ_k are learnable parameters for k-th Gaussian mixtures. Figure 2b shows how our Gaussian mixture functions allows to model fine-grained details as well as successfully reconstruct hair.

Integrated SDF with Cone Cast. However, unlike previous approaches that solely perform volume rendering, we have identified inconsistencies and visible artifacts on the geometry during hybrid rendering, as can be observed in the bottom-right visualization of Figure 5. These artefacts result from inaccuracies in estimating volume densities near the outer surface. Inspired by mip-NeRF (Barron et al. 2021), we propose to consider the ray as a cone rather than a single infinitesimally narrow line within the volume rendering region. To this aim, we investigate the evaluation of the SDF expected value within a conical frustum (a section of the cone) for each sample.

As Figure 2a shows, the apex of that cone lies at $\mathbf{o}$ with the axis along the viewing ray. The radius of the cone at location $\mathbf{o} + t \cdot \mathbf{d}$ is r. The set of positions $\mathbf{x}$ within a conical frustum between $[t - h, t + h]$ is:

$$F(\mathbf{x}, \mathbf{o}, \mathbf{d}, r, t, h) = \mathbb{1}\left\{ (t - h < \mathbf{d}^T(\mathbf{x} - \mathbf{o}) < t + h) \right.$$
$$\left. \wedge \left(\frac{\mathbf{d}^T(\mathbf{x} - \mathbf{o})}{\|\mathbf{x} - \mathbf{o}\|_2} > \frac{t}{\sqrt{t^2 + r^2}} \right) \right\}, \qquad (7)$$

where $\mathbb{1}$ is an indicator function: $F(\mathbf{x}, \cdot)$ iff $\mathbf{x}$ is within the conical frustum defined by $(\mathbf{o}, \mathbf{d}, r, t, h)$.

Theorem 1. *Inside a uniform signed distance field, where gradient directions on each location are parallel, given a known SDF value s_c at location $\mathbf{x}_c$ and gradient direction $\mathbf{n}_c$, the SDF value s at any location $\mathbf{x}$ is computed as*

$$s = \Phi(\mathbf{x}) = s_c + \mathbf{n}_c \cdot (\mathbf{x}_c - \mathbf{x}). \qquad (8)$$

The proof of this theorem is provided in Appendix. By assumption of the uniform SDF within each conical frustum, we compute the expected SDF values following Theorem 1:

$$E[s] = \frac{\int \Phi(\mathbf{x}) \cdot F(\mathbf{x}, \mathbf{o}, \mathbf{d}, r, t, h) \, d\mathbf{x}}{\int F(\mathbf{x}, \mathbf{o}, \mathbf{d}, r, t, h) \, d\mathbf{x}}. \qquad (9)$$

Here we compute both $E[s]$ and $E[s^2]$, for the details, refer to the Appendix,

$$E[s] = s_c + \frac{2h}{3t} \cdot \mathbf{d}^T \mathbf{n}_c,$$
$$E[s^2] = s_c^2 + \frac{t^2 h^2 \cdot (\mathbf{d}^T \mathbf{n}_c)^2 + 4t \cdot h^2 s_c \cdot \mathbf{d}^T \mathbf{n}_c}{3t^2 + h^2}, \qquad (10)$$

and thus,

$$\mathrm{E}[(\Phi(\mathbf{x}) - \mu_k)^2] = \mathrm{E}[s^2] - 2 \cdot \mu_k \mathrm{E}[s] + \mu_k^2. \quad (11)$$

Replacing the SDF value with the expectation in Eq. 6 gives the integrated Gaussian mixture SDF-to-density function:

$$\sigma(\mathbf{x}) = \sum_{k=0}^{K} \alpha_k \cdot \exp(-\frac{\mathrm{E}[(\Phi(\mathbf{x}) - \mu_k)^2]}{\beta_k^2}), \quad (12)$$

Note the proposed integrated SDF can also be used in other SDF-to-density functions (Yariv et al. 2021; Wang et al. 2021) $\Psi(\mathrm{E}[\Phi(\mathbf{x})])$.

Training

Our framework consists of three modules with learnable parameters: i) the SDF network $\Phi(\mathbf{x})$, ii) the implicit renderer $\mathbf{C} = M(\hat{\mathbf{x}}, \hat{\mathbf{n}}, \mathbf{d}, f)$, and iii) the SDF-to-density function $\Psi(s)$. We train our network by randomly sampling a set of pixels on each training image and minimize the sum of the overall loss computed on each pixel:

$$\mathcal{L} = \mathcal{L}_{rgb} + w_{mask}\mathcal{L}_{mask} + w_E\mathcal{L}_E + w_{Sp}\mathcal{L}_{Sp}, \quad (13)$$

where w_{mask}, w_E, and w_{Sp} are weights that balance each loss term.

Mask Loss $\mathcal{L}_{mask}$. We use a state-of-the-art human matting technique (Lin et al. 2021) to extract a saliency of the foreground instance, which provides an object likelihood per pixel. We threshold the saliency with $thr_s = 0.002$ and $thr_h = 0.9$ to generate the mask ground truth O_h and O_s for both SDF $\Phi_h(\mathbf{x})$ and $\Phi_s(\mathbf{x})$. Then a binary cross-entropy (BCE) loss is used to train the SDF network,

$$\mathcal{L}_{mask} = \mathrm{BCE}(\Phi_h(\hat{\mathbf{x}}), O_h) + \mathrm{BCE}(\Phi_s(\hat{\mathbf{x}}), O_s). \quad (14)$$

Photometric Loss $\mathcal{L}_{rgb}$. We compute the L1 loss between the constructed RGB value and the ground truth value:

$$\mathcal{L}_{rgb} = \sum_{p\in\mathcal{M}} \|C_p - \hat{C}_p\|, \quad (15)$$

where $\mathcal{M}$ is the mask of foreground regions, $\hat{C}_p$ is the ground-turth color on pixels p.

Specularity Loss $\mathcal{L}_{Sp}$. We introduce a new loss to regularize the spurious specularity appearing during novel view synthesis, which could cause color drifting to white or black. This loss penalizes sudden changes in the derivatives of appearance w.r.t. viewing direction, that is,

$$\mathcal{L}_{Sp} = \|\frac{\partial M(\hat{\mathbf{x}}, \hat{\mathbf{n}}, \mathbf{d}, f)}{\partial \mathbf{d}}\|_2. \quad (16)$$

Eikonal Loss $\mathcal{L}_E$. We use the Eikonal regularization on both SDFs (Gropp et al. 2020), *i.e.*,

$$\mathcal{L}_E = \mathrm{E}_{\mathbf{x}}[(\|\nabla\Phi_h(\mathbf{x})\| - 1)^2] + \mathrm{E}_{\mathbf{x}}[(\|\nabla\Phi_s(\mathbf{x})\| - 1)^2]. \quad (17)$$

where $\mathbf{x}$ is uniformly distributed inside the scene.

Additional Details. We follow IDR (Yariv et al. 2020) to pass through gradients toward the traced surface points. Note that different from IDR, in HISR, the differentiable surface points get gradients not only from the implicit renderer, but also from sampling locations in the volume rendering. We also improve the surface tracing process, for additional details refer to Appendix.

Experiments

Implementation Details

We implement our framework using PyTorch. Specifically, we use the Adam optimizer (Kingma and Ba 2014) with the learning rate $lr = 0.005$ and train for 6000 epochs per scene. At each training step, we randomly sample 7200 pixels with 80% inside the saliency. For volume rendering, we adaptively sample at most 400K points among all viewing rays in the translucent regions at each rendering step. We use a NVIDIA Tesla V100 GPU for training and it takes ~30h to train on each instance, and 110s to render each image in the original 2K resolution. As a comparison, NeRF takes 204s, IDR 68s, NeuS 380s and VolSDF 670s.

Datasets

We evaluate the effectiveness of our approach on three diverse datasets, encompassing real and synthetic human data, as well as a separate dataset for objects. This comprehensive evaluation allows us to showcase the broad applicability and versatility of our approach across different applications.

WIDC dataset contains sequences of real humans in motion captured with a 3dMD full-body scanner. The 3dMD scanner comprises 32 to 35 calibrated high-resolution RGB cameras (2048×2448) that capture a human in motion performing various actions and facial expressions and output a reconstructed 3D geometry and texture per frame. These scans can be noisy but capture facial expressions and fine-level details like cloth wrinkles. For each instance, we use 26 cameras for training, which focus on different body parts, and the other 6 to 9 cameras for evaluation, which capture the entire human. Examples of training images are provided in Appendix.

SynHuman dataset is generated by rendering a high-resolution animated 3D human model wearing synthetic clothes, including a simulated t-shirt and pants. Additionally, the dataset includes a hair groom with realistic hair strands, presenting a challenging test environment for our proposed approach. For training, we render images from 24 different cameras, while evaluation is conducted using images rendered from 6 distinct cameras for each instance. In total, we evaluate our approach on three instances within the dataset, allowing for a comprehensive assessment of its performance.

THUman dataset (Su et al. 2023) consists 128 viewing cameras from 4 scenes, with ~ 10 invalid cameras for each scene. Although we attempted to align the camera calibration as provided by the authors, we encountered an issue with slight displacement in the translation of the instance. This discrepancy may stem from the PyTorch3D camera system we utilized, which does not account for lens distortion in the capture system. To mitigate this, we optimized the translation during the training phase. Specifically, we employed an Adam optimizer to adjust the object's translation for each camera, using a learning rate of $1e - 4$. While we successfully trained our model on the THUman dataset, quantitative evaluations remain challenging due to the calibration issue. Also, due to the same reason, we are not able to provide comparisons with baselines on this dataset.

DTU dataset (Aanæs et al. 2016) comprises multi-view images (49 or 64 views) of various objects, captured using a

| Groundtruth | NeRF | IDR | VolSDF | NeuS | Ours |
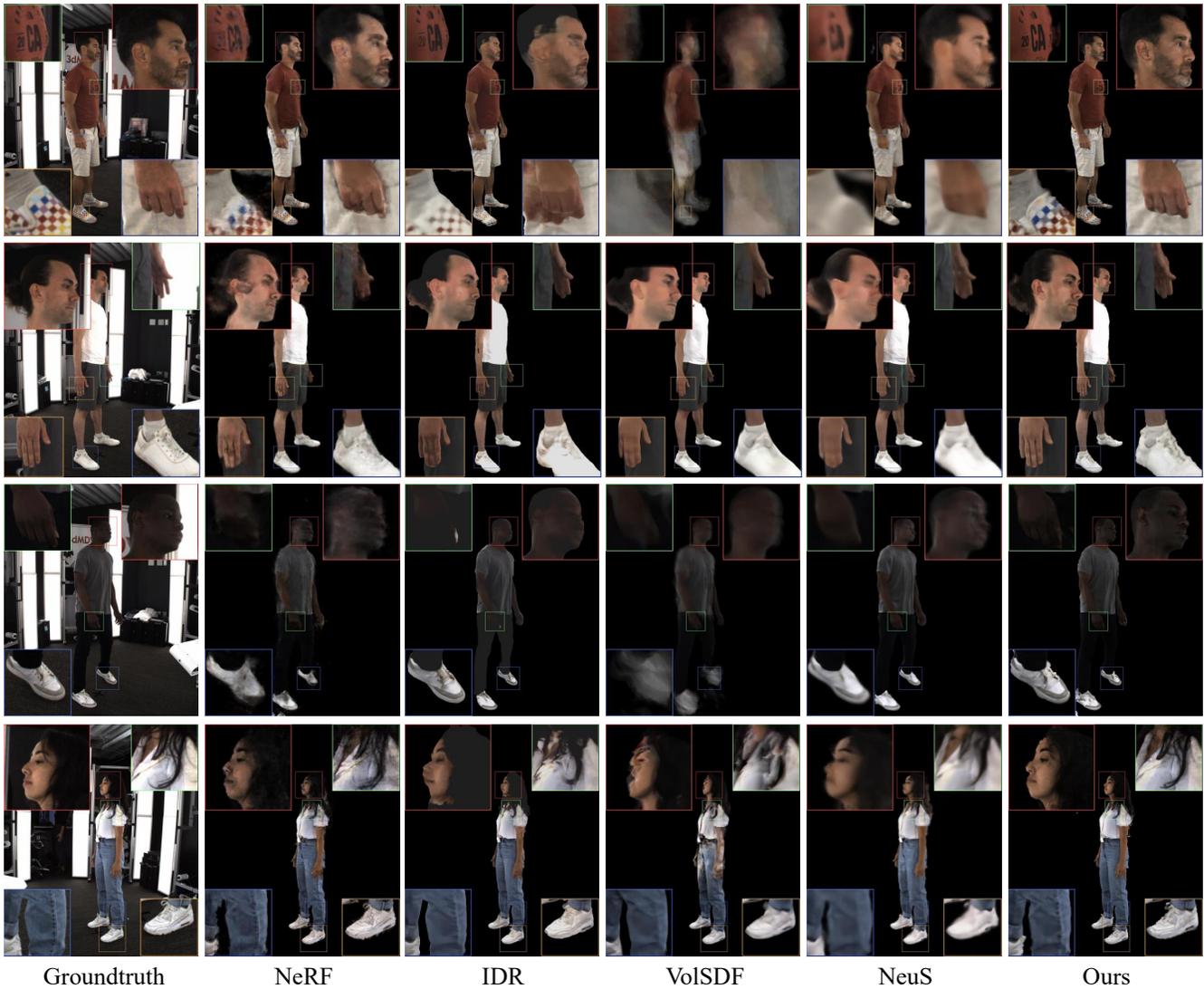
Figure 4: Qualitative comparisons with state-of-the-art approaches on novel view synthesis on WIDC dataset. To protect their privacy, we mask out part of people's faces.

PSNR ↑	S1.1	S1.2	S1.3	S2	S3	S4	S5.1	S5.2	S6	S7	S8.1	S8.2	Avg.
Nerf	28.62	31.65	29.62	24.72	29.33	32.67	26.42	26.76	26.06	23.39	21.21	27.47	27.33
IDR	27.98	30.73	31.37	24.84	29.33	30.01	26.40	26.39	26.74	27.74	28.55	28.85	28.24
VolSDF	28.01	30.37	30.28	24.52	25.94	29.79	25.75	26.28	21.92	23.44	31.14	30.51	27.33
Neus	29.05	31.91	32.30	26.17	30.76	32.90	27.54	27.48	27.77	28.86	31.88	32.19	29.90
GS	30.39	33.19	33.58	27.68	33.61	34.89	30.49	30.24	30.23	30.21	33.60	34.15	31.85
Ours	**31.51**	**34.41**	**34.07**	**28.57**	**33.76**	**35.14**	**30.62**	**31.34**	**30.51**	**31.16**	**34.49**	**34.48**	**32.49**

Table 2: Quantitative results and comparisons of PSNR between novel view synthesis and ground truth captures on WIDC dataset. Best scores are in bold.

light stage setup. This dataset provides a ground truth point cloud that serves as the basis for evaluating geometry reconstruction. In our work, we follow the settings established in previous research (Yariv et al. 2020). Specifically, we utilize a subset of 15 scenes from the dataset following previous

works.

Baselines and Evaluation Metrics

Baselines consist of NeRF (Mildenhall et al. 2020), IDR (Yariv et al. 2020), VolSDF (Yariv et al. 2021),
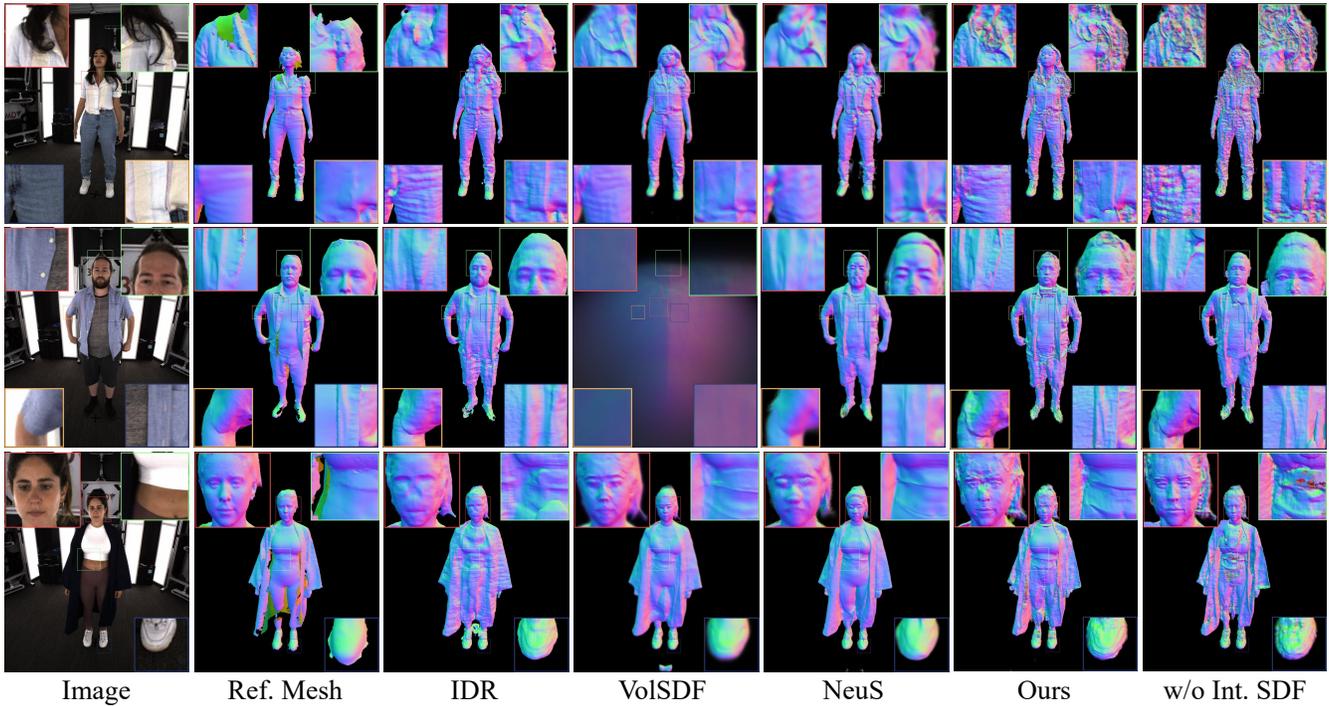
Figure 5: Qualitative comparisons with state-of-the-art approaches on geometry reconstruction. To show the details, we visualize both ours and baseline method with reconstructed surface normals. w/o Int. SDF is the ablation experiment for ours without the proposed integrated SDF.

CD* ↓	S1.1	S1.2	S1.3	S2	S3	S4	S5.1	S5.2	S6	S7	S8.1	S8.2	Avg.
Nerf	10.94	7.39	10.76	–	–	40.10	12.80	11.81	67.09	–	68.56	–	–
IDR	4.14	5.56	2.57	5.66	4.44	83.61	3.85	7.62	18.71	**3.68**	36.36	25.68	16.82
VolSDF	4.10	5.71	**1.05**	8.09	36.80	96.88	3.45	8.79	29.63	99.93	22.44	20.06	28.08
Neus	2.92	3.26	1.43	5.05	**2.43**	26.14	5.16	6.94	5.02	7.07	**11.09**	**14.60**	7.59
Ours	**1.39**	**2.06**	1.83	**4.81**	6.30	**7.87**	**3.31**	**6.16**	**3.86**	4.46	11.98	17.79	**5.99**

Table 3: Quantitative results and comparisons of Chamfer Distance (CD) between novel view synthesis and reference mesh reconstruction in WIDC dataset. Best scores are in bold.

NeuS (Wang et al. 2021), and Gaussian Splatting (GS) (Kerbl et al. 2023). For each baseline, we adapt the ray sampling and scene boundary to ensure the instance is placed inside the sampled range. Each baseline is trained for 6000 epochs on each scene. For NeuS and VolSDF, the mask loss proposed in NeuS is used during training. We observe NeRF and VolSDF may fail to converge and might generate empty images. If this happens, we retrain the approach until convergence and stop after three unsuccessful tries.

Evaluations are conducted for measurement of both novel view synthesis and reconstructed geometries. Specifically, we evaluate predicted novel views via PSNR, which computed via the L2 distance of groundtruth and reconstructed images. For the geometry, we first use the Marching Cubes algorithm (Lorensen and Cline 1987) to extract the surface mesh of the trained SDF. We use single-direction Chamfer Distance (CD*) for evaluation, which only computes the L1 distance from each vertex on ground truth mesh to the con-

	PSNR ↑	CD * ↓
with Gaussian Density	29.91	17.30
with Laplacian Density	29.76	30.86
w/o integrated SDF	30.28	20.41
w/o specularity loss	25.25	23.18
with Translucent region only	25.23	—
with Opaque region only	27.81	—
full model	**30.34**	**5.99**

Table 4: Ablation studies on WIDC dataset. Best scores are in bold.

structed one. This is because the captured 3D scans (*i.e.*, reference mesh) from 3dMD system may contain missing surfaces, as shown in Figure 5. For DTU datasets, we use the standard evaluation metrics PSNR and CD, following (Yariv et al. 2020). We also qualitatively evaluate the reconstructed
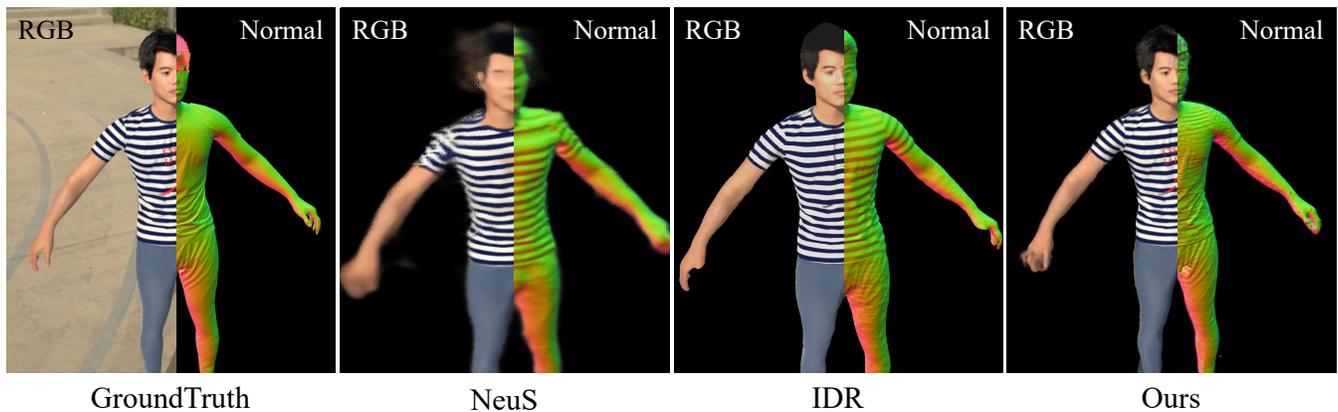
Figure 6: Qualitative comparisons for reconstructed geometry and novel view synthesis on SynHuman dataset
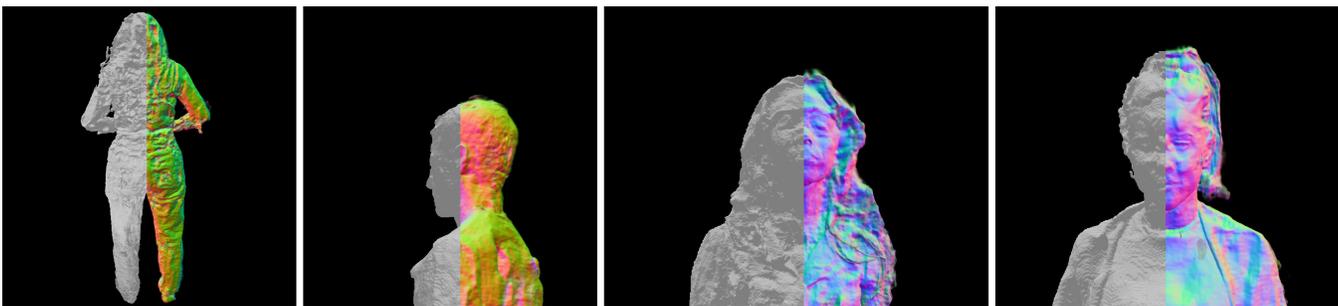
Figure 7: Comparison for reconstructed surface normals and reconstructed meshes. Specifically, we render the reconstructed meshes of Ours on the same pose of the rendered surface normals, and we concatenate them side-by-side.

geometry by visualizing the surface normals.

Quantitative Results and Qualitative Comparisons

WIDC. In Table 2 and Table 3, we provide a quantitative comparison between our method and various strong baselines. The results demonstrate that our approach outperforms all works in terms of both RGB reconstruction and reconstructed geometries. It is worth noting that the NeRF baseline may encounter difficulties in reconstructing via Marching Cubes if there are insufficient volumes with densities larger than the threshold value of $thr = 50$. In such cases, these entries in the table are denoted as "—". Figures 4 and 5 show the reconstructed geometries and novel view synthesis. The visualizations clearly showcase that our method excels in capturing fine-level geometric details compared to other reconstruction techniques. It is notable that our approach achieves better performance across the board and accurately reconstructs people with different hair styles and skin tones. Furthermore, our method exhibits a high level of rendering fidelity on the boundaries, surpassing approaches that solely rely on volume rendering.

SynHuman. In addition, we showcase the outcomes on the SynHuman dataset in Figure 6 and Table 6. The qualitative results highlight a substantial improvement in cloth reconstruction by our method compared to the baselines. Specifically, our approach yields more precise wrinkle details and lessens texture confusion with the geometries of the cloth.

These outcomes demonstrate the exceptional reconstruction quality attained through our methodology.

THUman. Qualitative results demonstrate our approach performs well in general 3D human reconstruction applications. From the results, we can observe we obtain high-fidelity reconstruction on both cloth and hair. Also, the results show our approach is feasible to achieve significantly better reconstruction quality with more multi-view input images, even with inaccurately calibrated cameras.

DTU. Table 5 and Figure 8 show the reconstruction results on the DTU dataset. The quantitative results illustrate significant improvements achieved by our method compared to the baseline method, *i.e.*IDR. Experiments on DTU dataset demonstrate that although HISR is specially designed for reconstruction of 3D humans, it can effectively perform reconstruction of generic objects. The visualization of surface normals shows ours more fine-grained details on the reconstructed geometry. As shown in the first column in Figure 8, "ours" achieves significantly better reconstruction of the geometry of the region under the scissors on the top.

Ablation Studies

In Table 4, we provide an ablation study to assess the individual contributions of each proposed module. In particular, we ablate *the Laplace function* and *the single Gaussian function* used in VolSDF and NeuS, which serve as replacements for our proposed Gaussian Mixture density. The results clearly
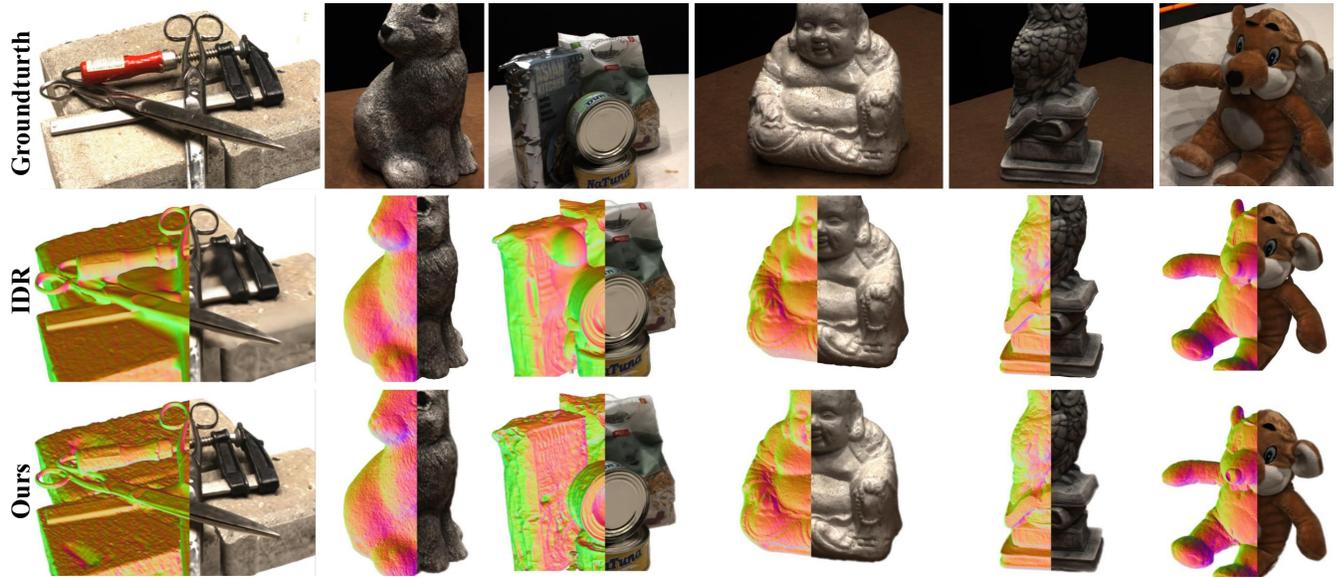
Figure 8: Qualitative comparisons for reconstructed geometry and novel view synthesis on DTU Dataset.

DTU	PSNR ↑		CD ↓	
Scene ID	IDR	Ours	IDR	Ours
24	23.55	24.39	1.24	0.90
37	21.08	22.53	1.22	0.56
40	24.62	25.66	0.87	1.08
55	23.64	24.07	0.39	0.33
63	25.47	26.47	0.54	0.39
65	23.26	26.55	0.82	0.54
69	22.36	24.31	0.42	0.49
83	21.97	24.42	2.22	2.17
97	23.16	24.18	0.92	0.51
105	22.97	26.73	0.99	1.00
106	22.17	26.04	0.51	0.84
110	23.07	23.86	0.97	1.36
114	25.04	25.81	0.25	0.43
118	24.18	26.25	0.54	0.49
122	27.42	27.30	0.63	0.44
Average	23.60	**25.24**	0.84	**0.77**

Table 5: Quantitative comparisons on DTU dataset.

	PSNR ↑	CD ↓
IDR	31.65	6.85
VolSDF	32.43	7.10
NeuS	33.21	10.58
Ours	**35.69**	**5.66**

Table 6: Quantitative comparisons on SynHuman dataset.

contribute to the final reconstruction.

Conclusion

In this paper, we introduced HISR, a novel hybrid implicit surface representation for photo-realistic 3D human reconstruction, employing a unique volume rendering scheme to maintain surface detail and realism. Our method computes the expected Signed Distance Function (SDF) values within a conical frustum, enhancing the quality of reconstructed geometry. Evaluated across various human and object reconstruction datasets, our approach surpasses baseline methods in both reconstructed geometry fidelity and novel view synthesis. This is due to our dual surface layer representation for opaque and translucent regions, allowing for nuanced rendering of complex human features like skin, hair, and clothing. In conclusion, our hybrid representation significantly advances neural reconstruction and rendering, particularly in handling heterogeneous geometries. Our approach achieves state-of-the-art results in 3D human reconstructions and shows promise in other objects.

indicate that these choices significantly reduce the quality of the reconstruction. Furthermore, we evaluate HISR *without the integrated SDF* module, by relying solely on the SDF value at the center of the conical frustum. For this ablation, we also show a qualitative comparison with our full approach in Figure 5. The results highlight the substantial contribution of the integrated SDF in achieving accurate reconstructed geometry. Then we perform an ablation study on the *specularity loss*, demonstrating its efficacy as a valuable regularizer that benefits both novel view synthesis and reconstruction tasks. Finally, we also report the novel view synthesis results when *only rendering translucent or opaque regions*, the results demonstrate both translucent or opaque are significantly

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision*, 1–16.

Alldieck, T.; Zanfir, M.; and Sminchisescu, C. 2022. Photo-realistic Monocular 3D Reconstruction of Humans Wearing Clothing. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *IEEE International Conference on Computer Vision*.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR*.

Cai, Y.; Wang, J.; Yuille, A.; Zhou, Z.; and Wang, A. 2023. Structure-Aware Sparse-View X-ray 3D Reconstruction. *arXiv preprint arXiv:2311.10959*.

Chen, X.; Zhang, Q.; Li, X.; Chen, Y.; Feng, Y.; Wang, X.; and Wang, J. 2022. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12943–12952.

Chen, Z.; and Zhang, H. 2019. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.

Gao, Y.; Cao, Y.-P.; and Shan, Y. 2023. SurfelNeRF: Neural Surfel Radiance Fields for Online Photorealistic Reconstruction of Indoor Scenes. *arXiv preprint arXiv:2304.08971*.

Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*.

Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; and Tung, T. 2020. ARCH: Animatable Reconstruction of Clothed Humans. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Işık, M.; Rünz, M.; Georgopoulos, M.; Khakhulin, T.; Starck, J.; Agapito, L.; and Nießner, M. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics (TOG)*, 42(4): 1–12.

Jiang, Y.; Hedman, P.; Mildenhall, B.; Xu, D.; Barron, J. T.; Wang, Z.; and Xue, T. 2022. AligNeRF: High-Fidelity Neural Radiance Fields via Alignment-Aware Training. *arXiv preprint arXiv:2211.09682*.

Kajiya, J. T.; and Herzen, B. P. V. 1984. Ray tracing volume densities. *ACM Transactions on Graphics*, 18(3): 165–174.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lin, S.; Yang, L.; Saleemi, I.; and Sengupta, S. 2021. Robust High-Resolution Video Matting with Temporal Guidance. *arXiv preprint arXiv:2108.11515*.

Lombardi, S.; Saragih, J.; Simon, T.; and Sheikh, Y. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 37(4): 1–13.

Lorensen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4): 163–169.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*.

Muller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4): 1–15.

Noguchi, A.; Sun, X.; Lin, S.; and Harada, T. 2021. Neural articulated radiance field. In *IEEE International Conference on Computer Vision*.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. In *IEEE International Conference on Computer Vision*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Love-grove, S. 2019a. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Love-grove, S. 2019b. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Pesavento, M.; Volino, M.; ; and Hilton, A. 2022. Super-resolution 3D Human Shape from a Single Low-Resolution Image. In *European Conference on Computer Vision*.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *IEEE International Conference on Computer Vision*.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. PI-FuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Su, Z.; Yu, T.; Wang, Y.; and Liu, Y. 2023. DeepCloth: Neural Garment Representation for Shape and Style Editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1581–1593.

Wang, A.; Kortylewski, A.; and Yuille, A. 2021. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. *arXiv preprint arXiv:2101.12378*.

Wang, A.; Ma, W.; Yuille, A.; and Kortylewski, A. 2023a. Neural Textured Deformable Meshes for Robust Analysis-by-Synthesis. *arXiv preprint arXiv:2306.00118*.

Wang, A.; Wang, P.; Sun, J.; Kortylewski, A.; and Yuille, A. 2022a. VoGE: a differentiable volume renderer using gaussian ellipsoids for analysis-by-synthesis. In *The Eleventh International Conference on Learning Representations*.

Wang, C.; Wang, A.; Li, J.; Yuille, A.; and Xie, C. 2023b. Benchmarking robustness in neural radiance fields. *arXiv preprint arXiv:2301.04075*.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Annual Conference on Neural Information Processing Systems*.

Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; et al. 2022b. Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. *arXiv preprint arXiv:2212.06135*.

Wang, Y.; Han, Q.; Habermann, M.; Daniilidis, K.; Theobalt, C.; and Liu, L. 2022c. NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction. *arXiv preprint arXiv:2212.05231*.

Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16210–16220.

Westover, L. 1990. Footprint evaluation for volume rendering. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, 367–376.

Xu, Q.; Xu, Z.; Philip, J.; Bi, S.; Shu, Z.; Sunkavalli, K.; and Neumann, U. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5438–5448.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume Rendering of Neural Implicit Surfaces. In *Annual Conference on Neural Information Processing Systems*.

Yariv, L.; Hedman, P.; Reiser, C.; Verbin, D.; Srinivasan, P. P.; Szeliski, R.; Barron, J. T.; and Mildenhall, B. 2023. BakedSDF: Meshing Neural SDFs for Real-Time View Synthesis. *arXiv preprint arXiv:2302.14859*.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *Annual Conference on Neural Information Processing Systems*.

Zwicker, M.; Pfister, H.; Van Baar, J.; and Gross, M. 2001. EWA volume splatting. In *IEEE Conference on Visualization*.