**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 \(||\nabla \Phi(x)|| - 1\), 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...
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.
The 3D object is represented as semi-transparent with a MLP, and a feature vector describing the geometry \( \hat{\mathbf{f}} \) at each location as:

\[
\hat{\mathbf{f}} = \mathbf{M}(\hat{\mathbf{x}}, \hat{\mathbf{n}}, \hat{\mathbf{d}}, f),
\]

where \( \hat{\mathbf{x}} = \hat{\mathbf{r}}(t) \) is the point location, \( \hat{\mathbf{n}} \) is the surface normal at \( \hat{\mathbf{x}} \), \( \mathbf{M} \) 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 \( \sigma(x) \) at each location in the scene \( x \in \mathbb{R}^3 \). 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 \( c_k \) and density \( \sigma_k \). 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):

\[
C = \sum_k T_k \left( 1 - \exp \left( -\sigma_k \left( t_{k+1} - t_k \right) \right) \right) c_k,
\]

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

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 \( \Phi \) is introduced:

\[
\sigma(x) = \Psi(-\Phi(x)),
\]

where \( \Psi \) 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 \( \sigma(x) \). 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 \( \Phi_h \) and soft SDF values \( \Phi_s \) at each location \( x \) in the space \( x \in \mathbb{R}^3 \) are:

\[
s_h = \Phi_h(x), \quad s_s = \Phi_s(x),
\]

where \( \Phi_h \) and \( \Phi_s \) 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 \( \sigma(x) = \Psi(-\Phi(x)) \).

**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 \( t_s \) and \( t_h \) such that \( \Phi_s(\mathbf{r}(t_s)) = 0 \) and

---

**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 \( t = t_s \) to \( t = t_h \).
Table 1: Possible cases in ray tracing.

<table>
<thead>
<tr>
<th>Case</th>
<th>Start</th>
<th>End</th>
<th>Final Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$C_{bg}$</td>
</tr>
<tr>
<td>2</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$C_h$</td>
</tr>
<tr>
<td>3</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$t_{s} = t_{h}$</td>
</tr>
<tr>
<td>4</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$t_{s} = t_{h}$</td>
</tr>
<tr>
<td>5</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$t_{s} = t_{h}$</td>
</tr>
</tbody>
</table>

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Figure 3: HISRC conducts volume rendering in translucent regions and surface rendering on hard surface. The numbers refer to the rendering cases presented in the table below.

The density of semi-transparent volumes is determined using the SDF-to-density function $\sigma(x) = \Psi(-\Phi(x))$. However, we have observed that the $\Psi$ 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:

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

where $\alpha_k$, $\beta_k$, and $\mu_k$ 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 $o$ with the axis along the viewing ray. The radius of the cone at location $o + t \cdot d$ is $r$. The set of positions $x$ within a conical frustum between $[t - h, t + h]$ is:

$$
F(x, o, d, r, t, h) = \{ (t - h < d^T (x - o) < t + h) \}
$$

$$
\wedge \left( \frac{d^T (x - o)}{|x - o|_2} > \frac{t}{\sqrt{r^2 + t^2}} \right) \}
$$

where $1$ is an indicator function: $F(x, o, d, r, t, h)$ is within the conical frustum defined by $(o, d, r, t, h)$.

**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 $x_c$ and gradient direction $n_c$, the SDF value $s$ at any location $x$ is computed as

$$
s = \Phi(x) = s_c + n_c \cdot (x_c - x).
$$

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:

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

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

$$
E[s] = \frac{2h}{3t} \cdot d^T n_c,
$$

$$
E[s^2] = \frac{h^2}{3t^2} \cdot d^T n_c^2 + \frac{4t \cdot h^2 s_c \cdot d^T n_c}{3t^2 + h^2}.
$$
and thus,
\[
E[(\Phi(x) - \mu_k)^2] = E[s^2] - 2 \cdot \mu_k E[s] + \mu_k^2.
\] (11)
Replacing the SDF value with the expectation in Eq. 6 gives the integrated Gaussian mixture SDF-to-density function:
\[
\sigma(x) = \sum_{k=0}^{K} \alpha_k \cdot \exp\left(-\frac{E[(\Phi(x) - \mu_k)^2]}{\beta_k^2}\right),
\] (12)
Note the proposed integrated SDF can also be used in other SDF-to-density functions (Yariv et al. 2021; Wang et al. 2021).

Training
Our framework consists of three modules with learnable parameters: i) the SDF network \( \Phi(x) \), ii) the implicit renderer \( C = M(x, n, d, f) \), 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:
\[
L = L_{rgb} + w_{mask} L_{mask} + w_E L_E + w_{Sp} L_{Sp},
\] (13)
where \( w_{mask}, w_E, \) and \( w_{Sp} \) are weights that balance each loss term.

Mask Loss \( L_{mask} \). 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 \( \Phi_h(x) \) and \( \Phi_s(x) \). Then a binary cross-entropy (BCE) loss is used to train the SDF network:
\[
L_{mask} = BCE(\Phi_h(\hat{x}), O_h) + BCE(\Phi_s(\hat{x}), O_s).
\] (14)

Photometric Loss \( L_{rgb} \). We compute the L1 loss between the constructed RGB value and the ground-truth value:
\[
L_{rgb} = \sum_{p \in M} \| C_p - \hat{C}_p \|,
\] (15)
where \( M \) is the mask of foreground regions, \( \hat{C}_p \) is the ground-truth color on pixels \( p \).

Specularity Loss \( L_{Sp} \). 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,
\[
L_{Sp} = \| \frac{\partial M(x, n, d, f)}{\partial d} \|_2.
\] (16)

Eikonal Loss \( L_E \). We use the Eikonal regularization on both SDFs (Gropp et al. 2020), i.e.,
\[
L_E = E_x[\| \nabla \Phi_h(x) \| - 1]^2] + E_x[\| \nabla \Phi_s(x) \| - 1]^2],
\] (17)
where \( x \) 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 \( \sim 30 \)h 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 VoIsDF 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 \( \times \) 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 \( \sim 10 \) 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
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.

<table>
<thead>
<tr>
<th>CD(\ast) ↓</th>
<th>S1.1</th>
<th>S1.2</th>
<th>S1.3</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5.1</th>
<th>S5.2</th>
<th>S6</th>
<th>S7</th>
<th>S8.1</th>
<th>S8.2</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nerf</td>
<td>10.94</td>
<td>7.39</td>
<td>10.76</td>
<td>--</td>
<td>40.10</td>
<td>12.80</td>
<td>11.81</td>
<td>67.09</td>
<td>--</td>
<td>68.56</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>IDR</td>
<td>4.14</td>
<td>5.56</td>
<td>2.57</td>
<td>5.66</td>
<td>4.44</td>
<td>83.61</td>
<td>3.85</td>
<td>7.62</td>
<td>18.71</td>
<td>3.68</td>
<td>36.36</td>
<td>25.68</td>
<td>16.82</td>
</tr>
<tr>
<td>VolSDF</td>
<td>4.10</td>
<td>5.71</td>
<td>1.05</td>
<td>0.89</td>
<td>36.80</td>
<td>96.88</td>
<td>3.45</td>
<td>8.79</td>
<td>29.63</td>
<td>99.93</td>
<td>32.44</td>
<td>20.06</td>
<td>28.08</td>
</tr>
<tr>
<td>Neus</td>
<td>2.92</td>
<td>3.26</td>
<td>1.43</td>
<td>5.05</td>
<td>2.43</td>
<td>26.14</td>
<td>5.16</td>
<td>6.94</td>
<td>7.02</td>
<td>7.07</td>
<td>11.09</td>
<td>14.60</td>
<td>7.59</td>
</tr>
<tr>
<td>Ours</td>
<td>1.39</td>
<td>2.06</td>
<td>1.83</td>
<td>4.81</td>
<td>6.30</td>
<td>7.87</td>
<td>3.31</td>
<td>6.16</td>
<td>3.86</td>
<td>4.46</td>
<td>11.98</td>
<td>17.79</td>
<td>5.99</td>
</tr>
</tbody>
</table>

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\(\ast\)) for evaluation, which only computes the L1 distance from each vertex on ground truth mesh to the con-

<table>
<thead>
<tr>
<th></th>
<th>PSNR ↑</th>
<th>CD(\ast) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Gaussian Density</td>
<td>29.91</td>
<td>17.30</td>
</tr>
<tr>
<td>w/o Laplacian Density</td>
<td>29.76</td>
<td>30.86</td>
</tr>
<tr>
<td>w/o integrated SDF</td>
<td>30.28</td>
<td>20.41</td>
</tr>
<tr>
<td>w/o specularity loss</td>
<td>25.25</td>
<td>23.18</td>
</tr>
<tr>
<td>w/o Translucent region only</td>
<td>25.23</td>
<td></td>
</tr>
<tr>
<td>w/o Opaque region only</td>
<td>27.81</td>
<td></td>
</tr>
<tr>
<td>full model</td>
<td><strong>30.34</strong></td>
<td><strong>5.99</strong></td>
</tr>
</tbody>
</table>

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

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. 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 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.
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 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.

**Acknowledgements.** We would like to thank Ziyan Wang for the help of running comparison experiments and tuning WIDC camera calibration.

---

**Table 5: Quantitative comparisons on DTU dataset.**

<table>
<thead>
<tr>
<th>Scene ID</th>
<th>IDR PSNR</th>
<th>IDR CD</th>
<th>Ours PSNR</th>
<th>Ours CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>23.55</td>
<td>1.24</td>
<td>24.39</td>
<td>0.90</td>
</tr>
<tr>
<td>37</td>
<td>21.08</td>
<td>1.22</td>
<td>22.53</td>
<td>0.56</td>
</tr>
<tr>
<td>40</td>
<td>24.62</td>
<td>0.87</td>
<td>25.66</td>
<td>1.08</td>
</tr>
<tr>
<td>55</td>
<td>24.64</td>
<td>0.39</td>
<td>24.07</td>
<td>0.33</td>
</tr>
<tr>
<td>63</td>
<td>25.47</td>
<td>0.54</td>
<td>26.47</td>
<td>0.39</td>
</tr>
<tr>
<td>65</td>
<td>23.26</td>
<td>0.82</td>
<td>26.55</td>
<td>0.54</td>
</tr>
<tr>
<td>69</td>
<td>22.36</td>
<td>0.42</td>
<td>24.31</td>
<td>0.49</td>
</tr>
<tr>
<td>83</td>
<td>21.97</td>
<td>2.22</td>
<td>24.42</td>
<td>2.17</td>
</tr>
<tr>
<td>97</td>
<td>23.16</td>
<td>0.92</td>
<td>24.18</td>
<td>0.51</td>
</tr>
<tr>
<td>105</td>
<td>22.97</td>
<td>0.99</td>
<td>26.73</td>
<td>1.00</td>
</tr>
<tr>
<td>106</td>
<td>22.17</td>
<td>0.51</td>
<td>26.04</td>
<td>0.84</td>
</tr>
<tr>
<td>110</td>
<td>23.07</td>
<td>0.97</td>
<td>23.86</td>
<td>1.36</td>
</tr>
<tr>
<td>114</td>
<td>25.04</td>
<td>0.25</td>
<td>25.81</td>
<td>0.43</td>
</tr>
<tr>
<td>118</td>
<td>24.18</td>
<td>0.54</td>
<td>26.25</td>
<td>0.49</td>
</tr>
<tr>
<td>122</td>
<td>27.42</td>
<td>0.63</td>
<td>27.30</td>
<td>0.44</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>23.60</strong></td>
<td><strong>0.84</strong></td>
<td><strong>25.24</strong></td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>

**Table 6: Quantitative comparisons on SynHuman dataset.**

<table>
<thead>
<tr>
<th></th>
<th>PSNR ↑</th>
<th>CD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDR</td>
<td>31.65</td>
<td>6.85</td>
</tr>
<tr>
<td>VoISDF</td>
<td>32.43</td>
<td>7.10</td>
</tr>
<tr>
<td>NeuS</td>
<td>33.21</td>
<td>10.58</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>35.69</strong></td>
<td><strong>5.66</strong></td>
</tr>
</tbody>
</table>

**Figure 8: Qualitative comparisons for reconstructed geometry and novel view synthesis on DTU Dataset.**
References


5308