2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.00029

Learning Visibility Field for Detailed 3D Human Reconstruction and Relighting

Ruichen Zheng*,[1,2], Peng Li*,[1], Haoqian Wang[1], Tao Yu[1]
[1]Tsinghua University, China [2]Weilan Tech, Beijing, China

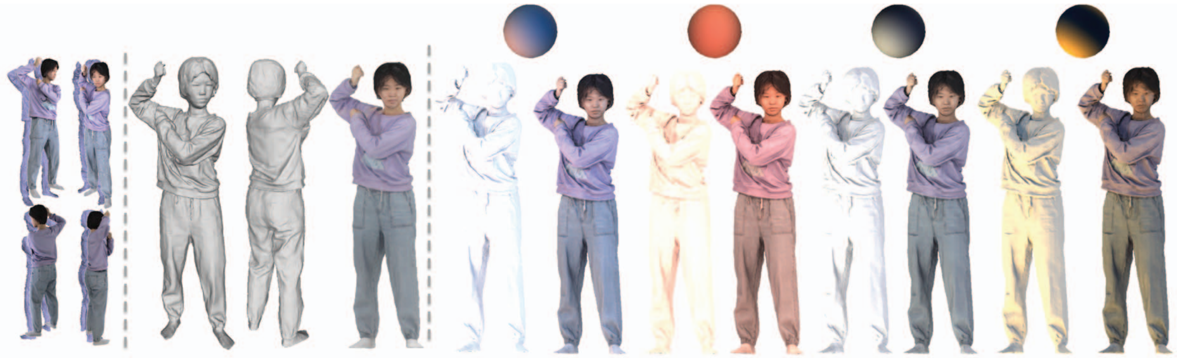


Figure 1. Given sparse-view RGBD input (left), our method reconstructs detailed geometry (middle left), albedo (middle right), and visibility for realistic 3D human relighting under different illuminations (right).

Abstract

Detailed 3D reconstruction and photo-realistic relighting of digital humans are essential for various applications. To this end, we propose a novel sparse-view 3d human reconstruction framework that closely incorporates the occupancy field and albedo field with an additional visibility field–it not only resolves occlusion ambiguity in multiview feature aggregation, but can also be used to evaluate light attenuation for self-shadowed relighting. To enhance its training viability and efficiency, we discretize visibility onto a fixed set of sample directions and supply it with coupled geometric 3D depth feature and local 2D image feature. We further propose a novel rendering-inspired loss, namely TransferLoss, to implicitly enforce the alignment between visibility and occupancy field, enabling end-to-end joint training. Results and extensive experiments demonstrate the effectiveness of the proposed method, as it surpasses state-of-the-art in terms of reconstruction accuracy while achieving comparably accurate relighting to ray-traced ground truth.

1. Introduction

3D reconstruction and relighting are of great importance in human digitization, especially in supporting realistic rendering in varying virtual environments, that can be widely applied in AR/VR [32, 37], holographic communication [24, 63], movie and gaming industry [7].

Traditional methods often require dense camera setups using multi-view stereo, non-rigid registration and texture mapping [9, 13]. To enhance capture realism, researchers have extended them with additional synchronous variable illumination systems, which aid photometric stereo for detail reconstruction and material acquisition [12]. However, these systems are often too complex, expensive and difficult to maintain, thus preventing widespread applications.

By leveraging deep prior and neural representation, sophisticated dense camera setups can be reduced to a single camera, leading to blossoms in learning-based human reconstruction. In particular, encoding human geometry and appearance as continuous fields using Multi-Layer Perceptron (MLP) has emerged as a promising lead. Starting from Siclope [36] and PIFu [43], a series of methods [15] improve the reconstruction performance in speed [11, 25], quality [44], robustness [65, 66] and light decoupling [1]. However, single-view reconstruction quality is restricted by its inherent depth ambiguity, thus limiting its application under view-consistent high-quality requirements.

Therefore, as the trade-off between view coverage and system accessibility, sparse-view reconstruction has become a research hotpot. The predominant practice is to project the query point onto each view to interpolate local features, which are then aggregated and fed to MLP for inference [6, 16, 43, 47, 52, 61]. This method suffers from occlusion ambiguity, where some views may well

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.00029

be occluded, and mixing their features with visible ones causes inefficient feature utilization, thus penalizing the reconstruction quality [43]. A natural solution is to filter features based on view visibility. Human templates such as SMPL [29] can serve as effective guidance [2,40,56,64], but introduce additional template alignment errors and therefore cannot guarantee complete occlusion awareness. Function4D [61] leverages the truncated Projective Signed Distance Function (PSDF) for visibility indication, but its level of details is susceptible to depth noise.

To this end, we directly model a continuous visibility field, which can be efficiently learned with our proposed framework and discretization technique using sparse-view RGB-D input. The visibility field enables efficient visibility query, which effectively guides multi-view feature aggregation for more accurate occupancy and albedo inference. Moreover, visibility can also be directly used for light attenuation evaluation–the key ingredient in achieving realistic self-shadowing. When supervising jointly with our novel TransferLoss, the alignment between the visibility field and occupancy field can be implicitly enforced without between-field constraints, such as matching visibility with occupancy ray integral. We train our framework end-to-end and demonstrate its effectiveness in detailed 3D human reconstruction by quantity and quality comparison with the state-of-the-art. We directly relight the reconstructed geometry with inferred visibility using diffuse Bidirectional Reflectance Distribution Function (BRDF) as in Fig. 1, which achieves photo-realistic self-shadowing without any post ray-tracing steps. To conclude, our contributions include:

- An end-to-end framework for sparse-view detailed 3D human reconstruction that also supports direct self-shadowed relighting.
- A novel method of visibility field learning, with the specifically designed TransferLoss significantly improves field alignment.
- A visibility-guided multi-view feature aggregation strategy that guarantees occlusion awareness.

2. Background and Related Work

Neural human synthesis The literature on human reconstruction is vast and rapidly growing. Here, we only review and contrast with closely related work and refer readers to surveys [55, 59] for comprehensive reviews.

Neural implicit representation encodes geometry as a function of spatial coordinate using MLP. This representation is appealing as it is naturally differentiable, has exceptional expressiveness yet maintains compact memory footprint. The pioneer work follows an encoder-decoder-like architecture, where the globally encoded feature is applied to condition spatial coordinates to infer low-level geometric details [5, 33, 34, 38]. This highly unbalanced information flow limits its capability to represent mere simple shapes [39]. PIFu [43] proposes to replace the global feature with pixel-aligned local features, which captures convolutional inductive bias and achieves highly detailed human reconstruction. It inspires a variety of works, ranging from quality improvement [44], parametric model extension [2, 65], animation support [14, 17], light estimation and relighting [1].

By averaging local features across views, the pixel-aligned framework can be easily extended to multi-view settings [43]. However, simple averaging diminishes high frequency details, yielding overly smoothed geometry. Moreover, it treats all views as equally visible even for partially occluded regions, resulting in inaccurate and erroneous reconstruction. Zhang *et al*. [64] addresses the occlusion ambiguity by incorporating the attention mechanism, which weights features by their learned cross-view correlations. Despite its promising performance, self-attention introduces substantial memory and computation overhead. Yu *et al*. [61] additionally leverages global depth information, namely PSDF, to annihilate the ambiguity but is sensitive to depth noise. In contrast, we use robustly learned visibility to weight per-view contribution, which handles occlusion in a physically plausible manner while being more memory efficient to compute.

Relighting boils down to substituting the incident environment radiance and re-evaluating the rendering equation [19]. For a scene of reflectors, the outgoing radiance reflected at the surface point $\boldsymbol{x}$ in direction $\boldsymbol{\omega}_o$ can be described as:

$$L(\boldsymbol{x}, \boldsymbol{\omega}_o) = \int_{\Omega^+} L(\boldsymbol{x}, \boldsymbol{\omega}_i)\rho(\boldsymbol{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o)V(\boldsymbol{x}, \boldsymbol{\omega}_i)(\boldsymbol{n}\cdot\boldsymbol{\omega}_i)d\boldsymbol{\omega}_i \tag{1}$$

where $\boldsymbol{n}$ is unit surface normal at $\boldsymbol{x}$, Ω^+ is hemisphere of possible directions $\boldsymbol{\omega}_i$, $L(\boldsymbol{x}, \boldsymbol{\omega}_i)$ is incident radiance arriving $\boldsymbol{x}$ along $\boldsymbol{\omega}_i$, directly from environment or indirectly reflected. ρ is the BRDF that models the surface reflectance. V is the visibility function that describes whether light $\boldsymbol{x}$ is attenuated along $\boldsymbol{\omega}_i$, As the main contributor to realistic self-shadowing, visibility evaluation requires ray tracing geometry over all sample directions $\boldsymbol{\omega}_i$ per fragment [18], which is expensive and usually omitted [45] or baked offline using Precomputed Radiance Transfer (PRT) [48]. PRT rewrites Eq. (1) with the following transfer function

$$T(\boldsymbol{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o) = \rho(\boldsymbol{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o)V(\boldsymbol{x}, \boldsymbol{\omega}_i)(\boldsymbol{n}\cdot\boldsymbol{\omega}_i) \tag{2}$$

which is independent of light and can be precomputed and projected on the Spherical Harmonics (SH) basis as coefficients [42] ahead of time to save rendering cost.

However, PRT does not ameliorate visibility calculation complexity and requires recompute after deformation. Its

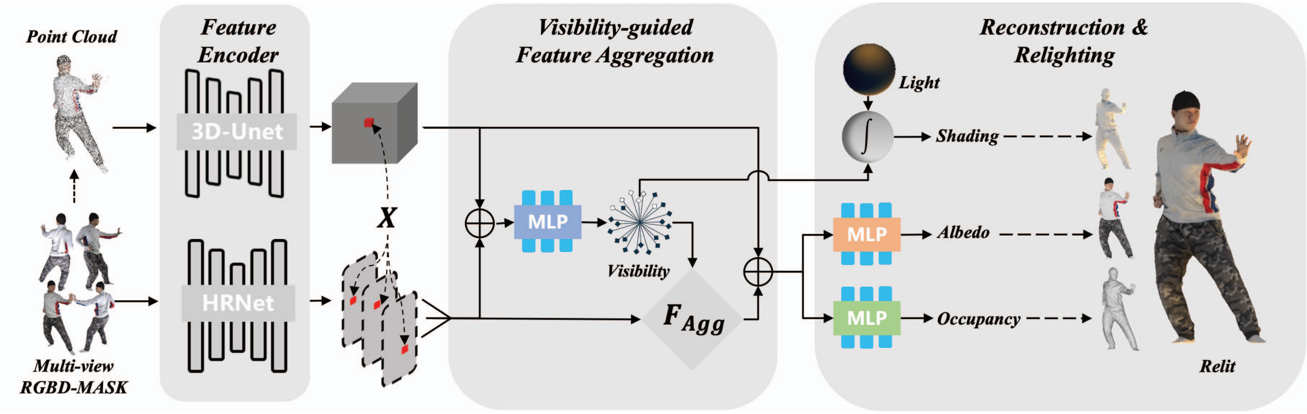


Figure 2. Method overview. Given sparse RGB-D frames as input, our framework first infer visibility, which is then applied to guide multi-view 2D feature aggregation for joint occupancy and albedo inference. Our framework is end-to-end trainable and produces high-fidelity human reconstruction that supports direct self-shadowed relighting without using any post ray-tracing steps.

limitation has been addressed by several learning-based approaches [20, 23, 26, 41, 53], which predict transfer coefficients using deep neural networks. Instead of its SH parameterization, we predict raw visibility, as it (1) can be used for feature aggregation, and (2) is also well defined in free space that can be learned without surface evaluation.

NeRF [35] models the surface density and employs volume rendering for high fidelity novel view synthesis. It has been extended with reflectance [3] and transfer function [31] to support relighting. Visibility is naturally defined in the density field, but its evaluation requires the integration of multiple density queries along the ray, which is equally inefficient and needs to be accelerated, such as using occlusion map [4] or MLP [50]. Although sharing similar ends, we achieve this with different means: (1) We emphasize accuracy by supervising with ground truth and use the TransferLoss to regularize the field alignment, rather than attempting to directly align the fields by matching inferred visibility with accumulated transmittance. (2) We extensively train our model on human scan dataset to ensure fast inference, rather than fitting it per-scene in order to render arbitrary within-scene objects.

3. Method

To introduce our method, we first define the visibility field and its discretization process, which prioritizes efficiency without compromising performance. We then outline our framework and visibility learning procedure. Finally, we showcase its application in sparse-view 3D human reconstruction, where it guides feature aggregation and enables direct self-shadowed relighting.

3.1. Visibility Field

For any point $\boldsymbol{X} \in \mathbb{R}^3$, whether it is visible $V \in \{0, 1\}$ along any view direction $\boldsymbol{\omega} \in \mathbb{R}^3$ of the unit sphere S forms a continuous field and can be parameterized using MLP:

$$\mathrm{MLP}_{\phi} : (\boldsymbol{X}, \boldsymbol{\omega}) \rightarrow V_{\phi} \in [0, 1],\ \boldsymbol{\omega} \in S \tag{3}$$

where ϕ is the MLP weights. Although Eq. (3) helps mitigate the cost of the ray integral, querying visibility over s directions still requires s calls per point of interest. This cost is exacerbated in pixel-aligned settings due to excessive point-wise feature $\boldsymbol{F}$ duplication as illustrated in Fig. 3a (top), leading to substantially high memory overhead.

We observe that, by uniformly sampling a discrete set of directions, visibility along any direction can be interpolated by Eq. (5) (Fig. 3c), with accuracy capped by the sample size. Thus, the $O(s)$ complexity can be reduced to a single MLP call plus an additional interpolation cost. To this end, we propose an alternative definition, by treating visibility as a function of $\boldsymbol{X}$ conditioned on a fixed set of n sample directions $\boldsymbol{\omega}^{(n)}$ as in Fig. 3a (bottom).

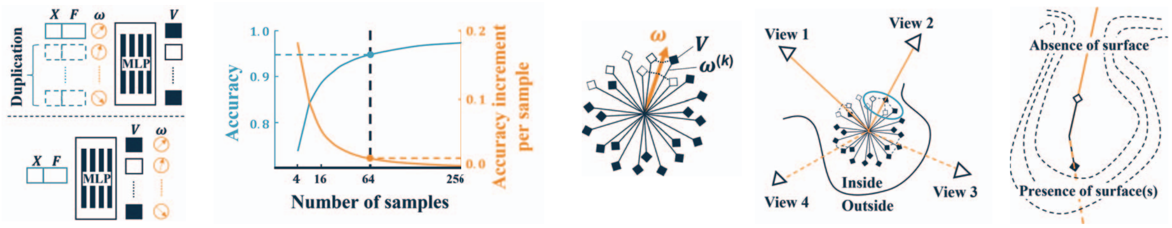
$$\mathrm{MLP}_{\phi|\boldsymbol{\omega}^{(n)}} : (\boldsymbol{X}, \boldsymbol{F}) \rightarrow V_{\phi|\boldsymbol{\omega}^{(n)}} \in [0, 1]^n, \boldsymbol{\omega}^{(n)} \in S \tag{4}$$

In practice, we interpolate by top k closest cosine distance with respect to $\boldsymbol{\omega}^{(n)}$:

$$\begin{gathered}\{\boldsymbol{\omega}^{(i)}\}_{i=1\ldots k} = topk(\boldsymbol{\omega}^{(n)} \cdot \boldsymbol{\omega}) \\ V(\boldsymbol{X}, \boldsymbol{\omega}) = \sum_{i=1}^{k} \frac{V_{\boldsymbol{\omega}^{(i)}}(\boldsymbol{X})(\boldsymbol{\omega}^{(i)} \cdot \boldsymbol{\omega})}{\sum_{j=1}^{k}(\boldsymbol{\omega}^{(j)} \cdot \boldsymbol{\omega})}\end{gathered} \tag{5}$$

The simplification in Eq. (4) leads to a surprisingly simple implementation–visibility prediction can be treated as $n - D$ binary classification supervised by BCE loss:

$$L_{\text{Vis}} = V_{\text{GT}} \cdot \log V_{\phi|\boldsymbol{\omega}^{(n)}} + (1 - V_{\text{GT}}) \cdot \log(1 - V_{\phi|\boldsymbol{\omega}^{(n)}}) \tag{6}$$



(a) (b) (c) (d) (e)

Figure 3. (a) Visibility as a continuous function of both point position $\boldsymbol{X}$ and sample direction $\boldsymbol{\omega}$ (top) leads to multiple queries and point-wise feature F duplication. In contrast, by fixing the sampling directions $\boldsymbol{\omega}^{(n)}$ (bottom), it only requires a single query. We extensively analyze our visibility field simplification strategy over 10% of training data and report average accuracy over the sample size n in (b) with the following procedure: For each sample size decision, we sample 10k query points per model (same sampling distribution as training data) with visibility evaluated over $\boldsymbol{\omega}^{(n)}$, that we apply cosine distance interpolation (c) to infer 10k randomly sampled validation directions per point. For aggregating multi-view features (d), we apply the same interpolation method as in (c) for each camera view direction to evaluate visibility along the back projection ray. (e) Directional visibility naturally constraints on the presence of surface.

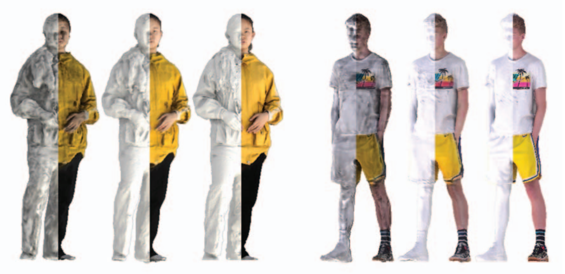


Figure 4. TransferLoss ablation: From left to right: w/o, w TransferLoss, ground truth. TransferLoss significantly mitigates rendering defects and improves near-surface visibility prediction accuracy (27.5% → 88.4% for synthesized testset, 20.8% → 84.7% for real-captured data).

3.2. Framework Overview

As shown in Fig. 2, our framework represents all fields using MLPs. Given sparse multi-view RGB-D frames $\{(\mathcal{I}_i, \mathcal{D}_i), i = 1 \ldots m\}$ with known camera projection matrices π_i, we first extract depth point cloud voxelized 3D feature and pixel-aligned RGB 2D feature to directly predict visibility. Then we utilize inferred visibility to guide the aggregation of multi-view 2D feature, which are paired with 3D feature for joint geometry and albedo inference.

3.3. Hybrid Feature Extraction

Conventionally, evaluating visibility requires tracing ray along the direction of interest and checking surface hits. In other words, it requires reasoning about the geometric feature of surfaces near that direction. Since we model both geometry (occupancy) and visibility, it is crucial to also bridge the two fields to incorporate their interconnection.

To this end, we adopt a hybrid feature extraction procedure, by separating depth from RGB and encoding it as 3D feature. Specifically, we follow [39] to unproject depth image into pointcloud, voxelize and filter using 3D convolutional as coarse feature volume. For a query point $\boldsymbol{X}$, we acquire its local 3D feature $F_{\mathcal{D}}$ by trilinear interpolation based on point coordinates and then share it with both visibility and occupancy MLPs. Intuitively, the 3D feature originates from voxelized point cloud, which can be viewed as noisy samples of the underlying surface (decision boundary of the occupancy field). By applying 3D convolution, the network reasons the coarse geometric surface feature over a sufficiently large receptive field, consequently aiding visibility inference. 3D feature is necessary for accurate visibility prediction as ablated in Fig. 8.

For multi-view RGB frames, we directly filter them using HRNet [51] to acquire 2D feature maps F_i. The local 2D features are then extracted in the pixel-aligned fashion as in [43], by projecting the point coordinate $\boldsymbol{X}$ onto each view as the image coordinate $\pi(\boldsymbol{X})$, then bilinearly interpolating the corresponding feature maps $F_i(\pi(\boldsymbol{X}))$. Compared to feature volume, 2D feature maps have much higher resolution and thus grant better details.

Follow Eq. (4), we infer $V_{\phi|\boldsymbol{\omega}^{(n)}}$ by providing 3D feature $F_{\mathcal{D}}$ and averaged 2D feature F_{avg}:

$$\text{MLP}_{\phi|\boldsymbol{\omega}^{(n)}} : (\boldsymbol{X}, F_{\mathcal{D}}, F_{avg}) \in [0,1]^n \tag{7}$$

3.4. Field Alignment Regularization

Given similar formulations, it is natural to train the visibility field and occupancy field together. However, as the point moves across the surface (occupancy classification boundary), its visibility changes drastically, ranging from fully occluded to partially visible. A slight misalignment between the two fields could cause inner visibility leakage and introduce conspicuous rendering defects as in Fig. 4. To enforce their alignment, a common practice is to explicitly constrain their correspondence, by matching visibility with surface queried along the ray [50], but at the expense of substantially large training overhead.

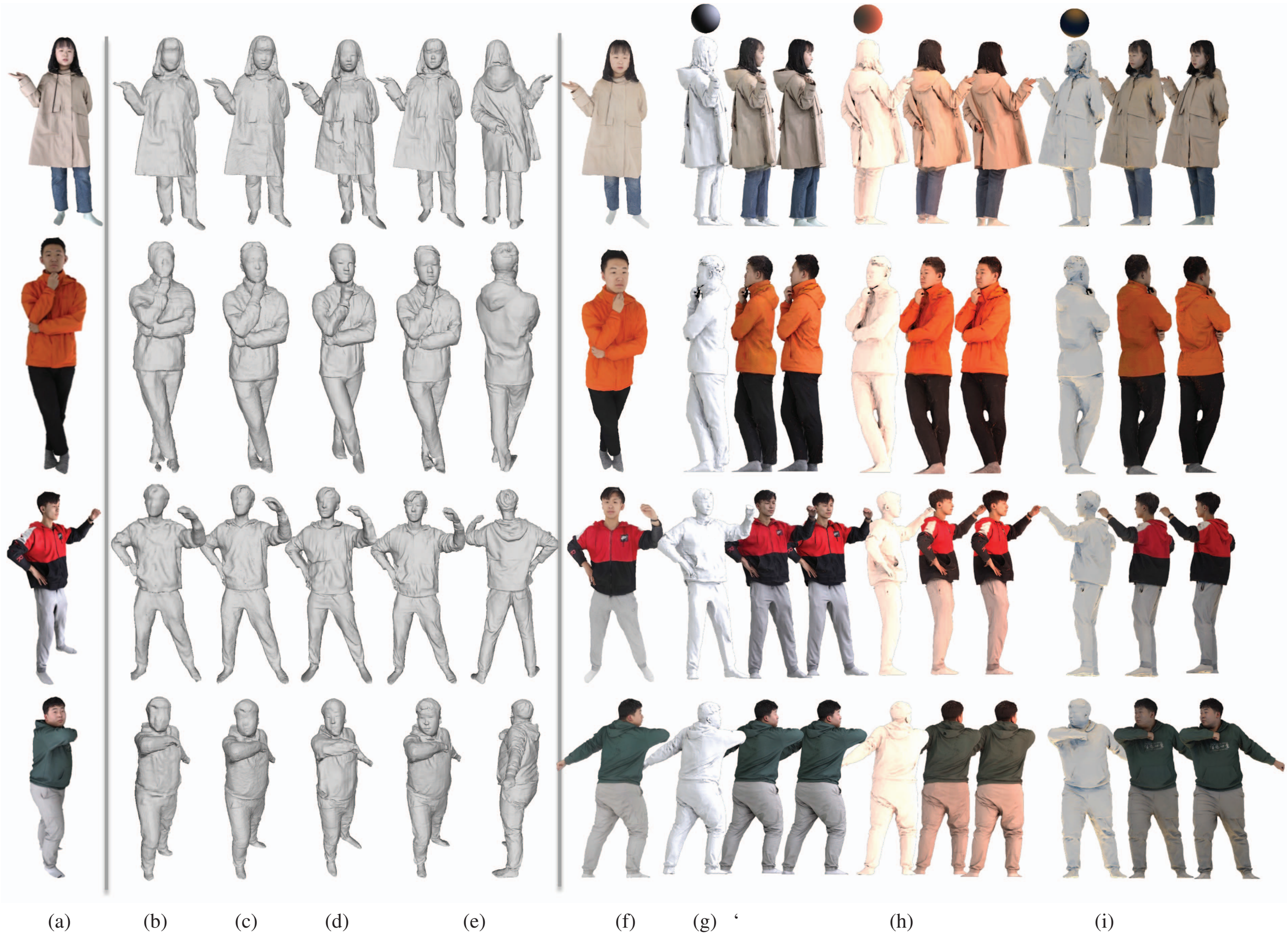


Figure 5. Qualitative comparisons on THUman2.0 [61]. We show (a) shaded color reference, geometry of (b) MV-PIFu [43], (c) Function4D [61], (d) MV-PIFuHD [44], (e) ours, (f) our albedo, (g-i) from left to right: our irradiance, our relighting and ground truth.

Since we bridge the two fields with the 3D feature reasoning about their correlation, we exploit the constraint of visibility over occupancy (Fig. 3e) by emphasizing the accuracy of near-surface visibility prediction, so that their alignment can be implicitly enforced. To this end, we propose a novel TransferLoss inspired by radiance transfer in Eq. (2):

$$L_{\text{Transfer}} = \sum_i |\hat{T}(\boldsymbol{X}, V_{\phi|\boldsymbol{\omega}^{(i)}}) - T(\boldsymbol{X}, V_{\text{GT}})| \qquad (8)$$

It supervises visibility, but (1) prioritizes the normal-facing directions in contrast to equal weighting in BCE (Eq. (6)) loss, and (2) follows the same parameterization for diffuse BRDF evaluation, which intuitively renders the light-independent part of the scene and perceptually penalizes the difference.

Since occupancy and visibility are jointly supervised, we directly sample query points on the ground truth mesh surface and use their normals to assist Eq. (8) evaluation for inferred visibility. TransferLoss effectively enforces fields alignment, resulting in significantly less rendering defects and higher relighting fidelity as ablated in Sec. 4.3 and Fig. 4.

3.5. Visibility-Guided Feature Aggregation

For each view, we trace back to its camera center and evaluate the directional visibility as described in Eq. (5). We then prioritize visible features over occluded ones using clamped negative log weighted average:

$$\begin{aligned} F_{agg} &= \sum_{i=1}^{m} W_{V_{\phi|\boldsymbol{\omega}^{(n)}}} F_i(\pi_i(\boldsymbol{X})) \\ W_{V_{\phi|\boldsymbol{\omega}^{(n)}}}(\boldsymbol{X}, \boldsymbol{\omega}) &= \max(-\log(1 - V_{\phi|\boldsymbol{\omega}^{(n)}}), 100) \end{aligned} \qquad (9)$$

Thus, occupancy and albedo are represented as:

$$\begin{aligned} \text{MLP}_{occ} &: (\boldsymbol{X}, F_{\mathcal{D}}, F_{agg}) \rightarrow [0, 1] \\ \text{MLP}_{albedo} &: (\boldsymbol{X}, F_{\mathcal{D}}, F_{agg}) \rightarrow \mathbb{R}^3 \end{aligned} \qquad (10)$$

3.6. Reconstruction and Relighting

During inference, we first apply visibility-guided aggregation to sample a grid of occupancy for surface mesh extraction [30]. We then pass the mesh vertices to obtain albedo and visibility using the same technique. Following Eq. (2), we compute per-vertex transfer coefficients and directly apply them for rasterized self-shadowed relighting.

4. Experiments

4.1. Experimental Settings

Training details. For training, we collect 400 high-quality clothed human scans from THUman2.0 [61], rotate each one around the yaw axis, and apply random shifts to obtain 60 views. For each view, we render 512×512 images of albedo, color using diffuse BRDF and depth fused with TOF depth sensors noise [10]. To simulate the incomplete depth caused by capture insensitivity for materials such as hair, we use [60] to mask hair out.

For each scanned mesh, we sample total 5k points for occupancy and visibility, with 4k near surface and 1k uniformly within the bounding volume. Near-surface points are sampled using normal distribution with standard deviation of 0.05, and we ensure that half of them with distances less than the standard deviation are used for albedo training. As described in Sec. 3.1 and Fig. 3b, we uniformly sample 64 fixed directions using the Fibonacci lattice and keep it consistent throughout the experiment. We use Embree [57] for ground truth visibility evaluation.

In addition to visibility, we supervise per-sample occupancy with BCE loss and albedo with L_1 loss. We also extract patches using depth-guided raymarching [27] and supervise its albedo using VGG perceptual loss.

We set the view number to 4 and train our framework using Adam [21] optimizer and Cyclic learning rate (lr) scheduler [49] over 600 epochs. The lr ranges from $5e{-}5$ to $5e{-}4$ every 5 epochs, with the max lr halved every 100 epochs.

Evaluation details. For evaluation, we prepare 100 training-excluded scans from THUman2.0, and an additional 100 scans from Twindom [46]. We follow the same rendering procedure as for training and report metrics averaged over the two sets. We further prepare real-captured RGB-D video sequences from a synchronized multi-Kinect capturing system that we leverage RVM [28] for foreground mask segmentation. For all experiments, we keep the number of views to 4 except for the view-number ablation. All meshes are extracted from 512^3 voxel using Marching Cube [30]. The relighting results are rendered using diffuse BRDF with inferred visibility and SH order of 2.

All experiments run on a PC with an Nvidia GeForce RTX3090 GPU and an Intel i7-8700k CPU. Our framework requires 50 ms for 3D and 100 ms for 2D feature extraction, and 200 ms per 2 million point queries for each of the 4-layer ResNet MLP decoders. The total reconstruction process takes approximately 3 seconds. After experimenting with TensorRT conversion, 3D and 2D feature extraction can be reduced to 10 and 7 ms, respectively. Since they can be performed in parallel, our framework has the potential to achieve real-time performance under heavy decoder optimization.

Methods	Metrics		
	NC ↑	CD (L_1) ↓	F-score (0.5%)↑
MV-PIFu [43]	0.912	0.145	0.624
MV-PIFuHD [44]	0.906	0.135	0.690
Function4D [61]	0.893	0.129	0.704
Ours	**0.917**	**0.122**	**0.736**

Table 1. Quantitative comparisons on reconstruction quality.

Metrics. We report normal consistency (NC) [33], L_1 Chamfer Distance (CD L_1) [8] and F-score [22] for geometric quality evaluation. Specifically, NC is calculated as the mean absolute dot product of normals of points sampled from reconstructed mesh and their closest neighbours' ones on ground truth mesh. We follow [33] to use 1/10 of the maximum bounding box edge as unit 1 for CD L_1 and 0.5% as the F-score distance threshold as suggested by [54].

To evaluate the quality of relighting, we adopt the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM) [58] and the learned perceptual image patch similarity (LPIPS) [62]. We mask out renderings with ground truth alpha channel and only report the average contributions of valid pixels.

4.2. Comparisons

Reconstruction. We compare our framework with the state-of-the-art prior-free sparse-view reconstruction approaches, namely multi-view PIFu (MV-PIFu) [43], multi-view PIFuHD (MV-PIFuHD) [44] and Function4D [61]. MV-PIFu takes RGB inputs and aggregates using averaging, which we adapt with additional depth input to ensure comparison consistency. PIFuHD integrates normal information, coarse-to-fine two-stage inference, and higher 1024×1024 resolution input. We self-implement multi-view RGB-D variants (MV-PIFuHD) with the same averaging aggregation as MV-PIFu. Normal maps are still inferred rather than being computed from depth maps due to noise concern. Function4D uses averaging in geometry inference as well, but its integration of the truncated PSDF serves as a strong signal to identify visible features and has been shown to generalize well on real captured data. For a fair comparison, we re-render and train all three approaches on our dataset until converge. Regrettably, we could only compare the geometry, since none of them estimates surface albedo.

Relighting. Limited by the dataset, which maps human into

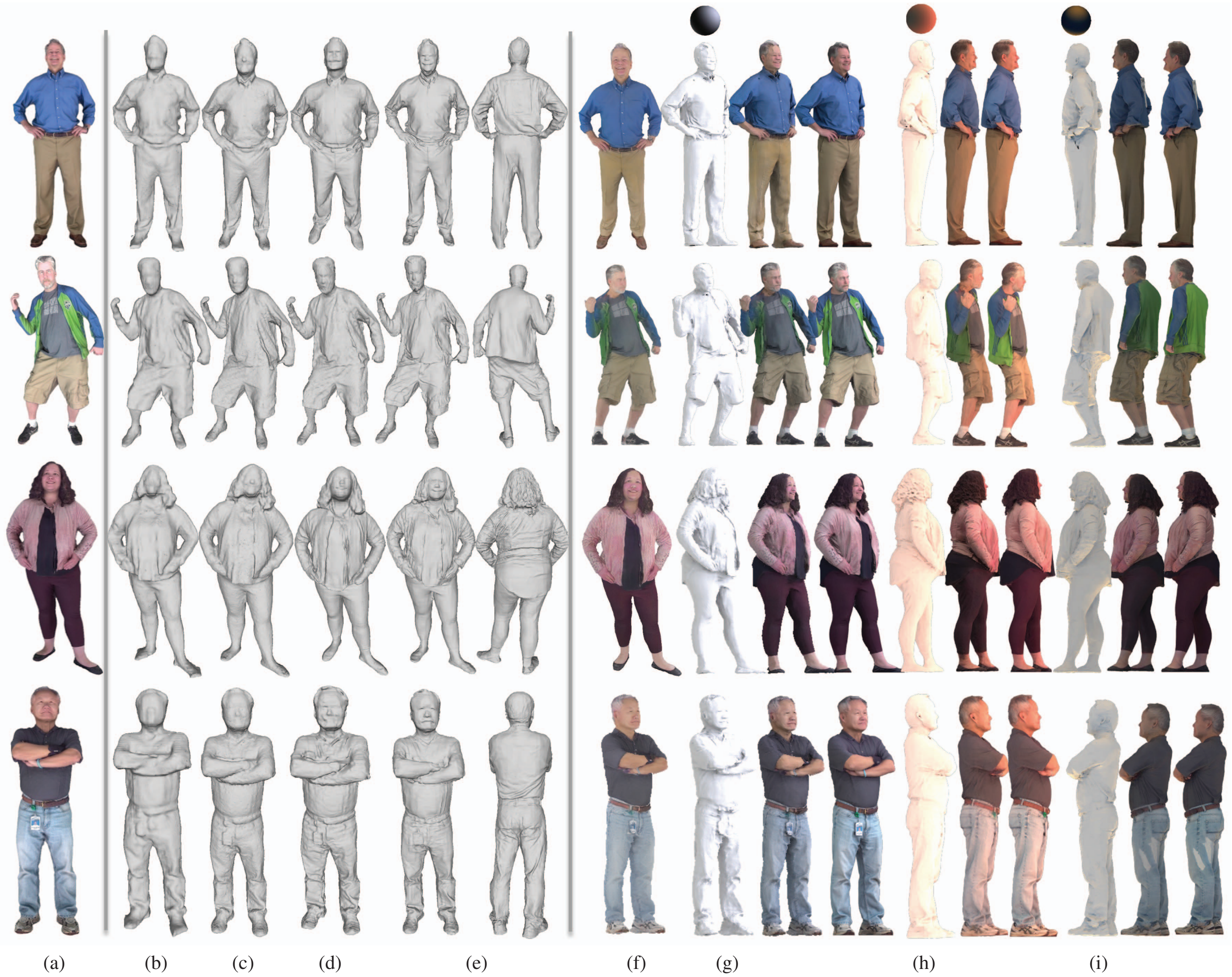


Figure 6. Qualitative comparisons on Twindom. From left to right: (a) shaded color reference, geometry of (b) MV-PIFu [43], (c) Function4D [61], (d) MV-PIFuHD [44], (e) ours, (f) our albedo, (g-i) from left to right: our irradiance, our relighting and ground truth.

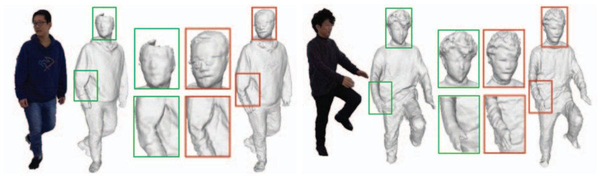


Figure 7. Visualization comparison between Function4D [61] (green) and our method (red) on data captured by Kinect.

single albedo texture without differentiating specular components and coefficients, we were only able to render using diffuse BRDF. This also limits our relighting comparison with NeRF-like method [3,4,50], where the view-dependent specular term is explicitly modeled and is crucial for rendering fidelity. Therefore, we directly compare our results to ray-traced ground truth to demonstrate our relighting performance.

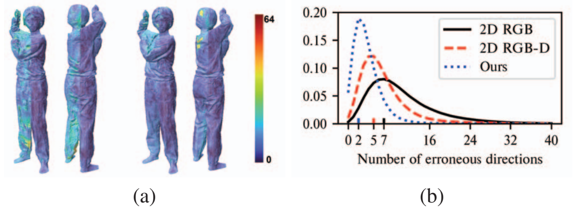


Figure 8. 3D Feature ablation. (a) We compare the visibility error maps using 2D RGB input (left half of first two), 2D RGB-D input (left half of last two) and ours (right half of all four). (b) We count per-vertex number of wrong directional visibility predictions (out of 64) and plot the normalized histogram.

Qualitative Comparison. Fig. 5 shows the reconstruction

Table 2. Feature aggregation ablation.

Method	Geometry Metric			Relit Rendering		
	NC↑	CD (L_1)↓	F-score (0.5%)↑	PSNR↑	SSIM↑	LPIPS↓
Average	0.911	0.129	0.696	17.933	0.650	0.327
Attention	0.909	0.127	0.710	19.148	0.713	0.259
Ours	**0.917**	**0.122**	**0.736**	**23.436**	**0.809**	**0.196**

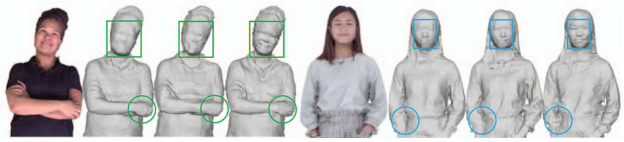


Figure 9. Feature aggregation ablation in geometry. From left to right: color reference, the results of average, attention and ours.

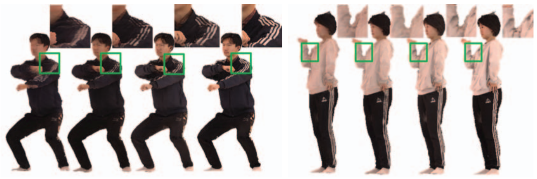


Figure 10. Feature aggregation ablation in rendering quality. From left to right: average, attention, and ours and ground truth.

on the synthetic datasets. Benefiting from visibility-guided feature aggregation, our method produces at least comparable details to MV-PIFuHD, but with simpler architecture, less input information and visual cues. Compared to oversmoothed geometry from Function4D, our hybrid features relax the dominant contribution of noisy depth input, leading to improved facial details. We further demonstrate our generalizability on real captured data in comparison with Function4D. As shown in Fig. 7, our method evidently produces more complete and detailed reconstructions, especially in regions of the eyes, glasses and hair.

Quantitative Comparison summarized in Tab. 1 is consistent with qualitative analysis, and our method outperforms others in all metrics.

4.3. Ablation Study

3D Feature. We ablate our hybrid feature with 2D RGB and 2D RGB-D variants. The results validate the necessity of the 3D feature, as it predicts visibility with higher accuracy (Fig. 8a) and lower error variance (Fig. 8b).

Visibility-guided Feature Aggregation. In comparison with other aggregation techniques, namely averaging [43] and self-attention [61], we implement them in our framework. Our strategy surpasses others in metrics (Tab. 2) and achieves sharper geometric details (Fig. 9), better rendering fidelity (Fig. 10), even near heavily folded clothing thanks to our occlusion-aware aggregation.

View number. Though we train our model with 4 views, it generalizes well across different view numbers and achieves similar performance under sufficient view coverage as ablated in Fig. 11. As view number increases, compared to degraded reconstruction quality from average and attention, our aggregation technique mitigates visibility ambiguity to leverage additional visual cues, resulting in better geometry.

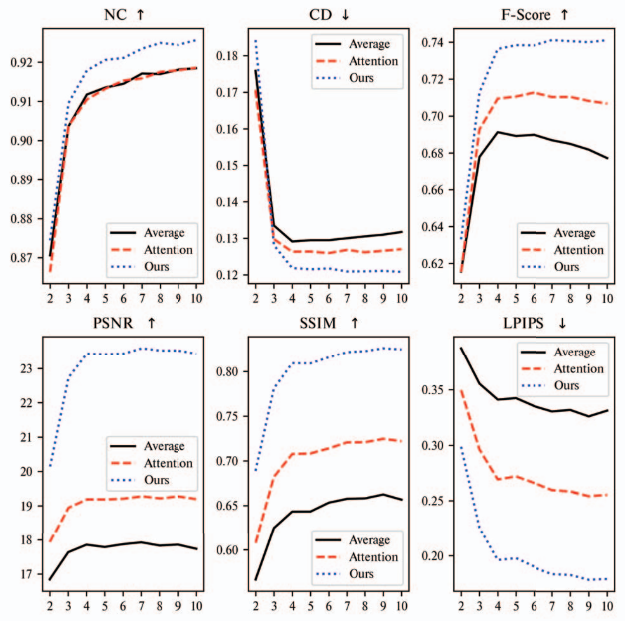


Figure 11. View number ablation.

TransferLoss. Finally, we verify the effectiveness of our proposed TransferLoss in Fig. 4. It enforces the alignment of the fields and significantly alleviates rendering defects.

5. Conclusion

Limitations. Our method cannot achieve real-time performance for interactive applications. We leave its acceleration using TensorRT and CUDA for future work. Moreover, our framework relies on depth input for accurate visibility prediction, it would be interesting to see if it can be extended to simpler RGB setup.

Conclusion. In this work, we integrate visibility into sparse-view reconstruction framework by exploiting its effective guidance in multi-view feature aggregation and direct support for self-shadowed relighting. Our discretization strategy and novel TransferLoss enable visibility to be learned jointly alongside occupancy in an end-to-end manner. Our paper demonstrates the effectiveness of visibility in simultaneous reconstruction and relighting and provides a good baseline for future work.

Acknowledgement. This work was supported by NSFC No.62171255, National Key R&D Program of China No.2022YFF0902201, Guoqiang Institute of Tsinghua University No.2021GQG0001 and Tsinghua-Toyota Joint Research Fund.

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 1, 2

[2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 2

[3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 3, 7

[4] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. 3, 7

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[6] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6981, 2020. 1

[7] Paul Debevec. Pursuing reality with image-based modeling, rendering, and lighting. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 1–16. Springer, 2000. 1

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 6

[9] Sean Ryan Fanello, Cem Keskin, Shahram Izadi, Pushmeet Kohli, David Kim, David Sweeney, Antonio Criminisi, Jamie Shotton, Sing Bing Kang, and Tim Paek. Learning to be a depth camera for close-range human capture and interaction. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 1

[10] Péter Fankhauser, Michael Bloesch, Diego Rodriguez, Ralf Kaestner, Marco Hutter, and Roland Siegwart. Kinect v2 for mobile robot navigation: Evaluation and modeling. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 388–394. IEEE, 2015. 6

[11] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: learning fourier occupancy field for monocular real-time human reconstruction. *arXiv preprint arXiv:2206.02194*, 2022. 1

[12] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1

[13] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015. 1

[14] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 2

[15] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 535–545, 2021. 1

[16] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 710–720. IEEE, 2021. 1

[17] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2

[18] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 388–405. Springer, 2022. 2

[19] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 2

[20] Yoshihiro Kanamori and Yuki Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *arXiv preprint arXiv:1908.02714*, 2019. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 6

[23] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. Single-image full-body human relighting. *arXiv preprint arXiv:2107.07259*, 2021. 3

[24] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. Project starline: A high-fidelity telepresence system. 2021. 1

[25] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision*, pages 49–67. Springer, 2020. 1

[26] Yue Li, Pablo Wiedemann, and Kenny Mitchell. Deep precomputed radiance transfer for deformable objects. *Pro-*

ceedings of the ACM on Computer Graphics and Interactive Techniques, 2(1):1–16, 2019. 3

[27] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 6

[28] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 6

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6

[31] Linjie Lyu, Ayush Tewari, Thomas Leimkühler, Marc Habermann, and Christian Theobalt. Neural radiance transfer fields for relightable novel-view synthesis with global illumination. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 153–169. Springer, 2022. 3

[32] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 1

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2, 6

[34] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Deep level sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802*, 2019. 2

[35] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[36] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019. 1

[37] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pages 741–754, 2016. 1

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[39] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 2, 4

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2

[41] Gilles Rainer, Adrien Bousseau, Tobias Ritschel, and George Drettakis. Neural precomputed radiance transfer. In *Computer Graphics Forum*, volume 41, pages 365–378. Wiley Online Library, 2022. 3

[42] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 2

[43] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 4, 5, 6, 7, 8

[44] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1, 2, 5, 6, 7

[45] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 2

[46] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15872–15882, 2022. 6

[47] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. *arXiv preprint arXiv:2207.08000*, 2022. 1

[48] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 527–536, 2002. 2

[49] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. 6

[50] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting

and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 3, 4, 7

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 4

[52] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6226–6237, 2021. 1

[53] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, volume 40, pages 205–216. Wiley Online Library, 2021. 3

[54] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414, 2019. 6

[55] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 2

[56] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[57] Ingo Wald, Sven Woop, Carsten Benthin, Gregory S Johnson, and Manfred Ernst. Embree: a kernel framework for efficient cpu ray tracing. *ACM Transactions on Graphics (TOG)*, 33(4):1–8, 2014. 6

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[59] 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*, volume 41, pages 641–676. Wiley Online Library, 2022. 2

[60] YBIGTA. Hair-segmentation. `https://pytorchhair.gitbook.io/project/`. 6

[61] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5746–5756, 2021. 1, 2, 5, 6, 7, 8

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[63] Yizhong Zhang, Jiaolong Yang, Zhen Liu, Ruicheng Wang, Guojun Chen, Xin Tong, and Baining Guo. Virtualcube: An immersive 3d video communication system. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2146–2156, 2022. 1

[64] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 2

[65] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 1, 2

[66] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 1