Millimetric Human Surface Capture in Minutes

Briac Toussaint, Laurence Boissieux, Diego Thomas, Edmond Boyer, Franco Jean-Sébastien

Millimetric Human Surface Capture in Minutes

BRIAC TOUSSAINT, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LJK, France
LAURENCE BOISSIEUX, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LJK, France
DIEGO THOMAS, Kyushu University, Japan
EDMOND BOYER, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LJK, France
JEAN-SÉBASTIEN FRANCO, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LJK, France

Fig. 1. Left: Reconstructions for 3 examples of dynamic human capture [Armando et al. 2023], including body and long hair motion. Right: illustration of results on the proposed MVMannequin scanned mannequin database: input view, hand-scanned reference, our reconstruction. Note in both cases, the fine captured surface detail such as hair strands, facial features, cloth folds and creases, and the sewing micro-geometry.

Detailed human surface capture from multiple images is an essential component for many 3D production, analysis and transmission tasks. Yet producing millimetric precision 3D models in practical time, and actually verifying their 3D accuracy in a real-world capture context, remain key challenges due to the lack of specific methods and data for these goals. We propose two complementary contributions to this end. The first one is a highly scalable neural surface radiance field approach able to achieve millimetric *precision* by construction, while demonstrating high compute and memory efficiency. The second one is a novel dataset, MVMannequin, of clothed mannequin geometry captured with a high resolution hand-held 3D scanner paired with calibrated multi-view images, that allows to verify the millimetric *accuracy* claim. Although our approach can produce such highly dense and precise geometry, we show how aggressive sparsification and optimizations of the neural surface pipeline allow estimations in minutes of computation time using only a few GB of GPU memory, while allowing for real-time millisecond neural rendering. On the basis of our framework and dataset, we show that our method achieves submillimetric accuracy and completeness for 77% of the points in less than 3 minutes of training time, with 68 viewpoints.

* Institute of Engineering Univ. Grenoble Alpes.

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12.
https://doi.org/10.1145/3680528.3687690

CCS Concepts: • **Computing methodologies** → **Computer vision representations**; **Reconstruction**; **Volumetric models**; **Rendering**.

Additional Key Words and Phrases: Neural radiance fields, Human surface capture, Differential rendering

ACM Reference Format:
Briac Toussaint, Laurence Boissieux, Diego Thomas, Edmond Boyer, and Jean-Sébastien Franco. 2024. Millimetric Human Surface Capture in Minutes. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3680528.3687690

1 INTRODUCTION

Passively and effectively capturing 3D geometry of humans in motion is a long pursued goal with many challenges. It is relevant to a large number of applications that benefit from accurate 3D human surface metrology, such as 3D digital production, virtual reality, telepresence, performance capture, medical monitoring, or sport gesture analysis. Multi-camera platforms have long been used for 3D capture tasks in human scenes, but have seldom been demonstrated to achieve millimetric precision and accuracy with practical resource usage in this context. This is due both to methodological challenges, and the lack of densely annotated real-world data of human bodies with fine surface details, such as facial features, cloth folds and creases, allowing to measure the 3D error.

Multi-view stereo [Furukawa and Ponce 2010; Schönberger and Frahm 2016] has long been the metrological basis for precise human

surface capture. Yet a stream of recent methods show that neural radiance approaches are promising for this task [Işık et al. 2023]. Originally relying on implicit opacity field scene representation and differential ray casting for rendering [Mildenhall et al. 2021], parameter-efficient overhauls propose explicit parametric Gaussians for volumetric opacity encoding and splatted rendering [Kerbl et al. 2023]. Initially focused on rendering applications, some promising variants also target surface extraction by regularizing volumetric opacities, *e.g.* [Huang et al. 2024], but have yet to surpass specialized surface-based methods for 3D accuracy and compute metrics.

With the objective to reach millimeter precision, we focus on neural methods with dedicated signed distance function surface parametrization [Wang et al. 2021; Yariv et al. 2020], as they yield state of the art surface reconstruction performance on generic object benchmarks such as [Aanæs et al. 2016; Yao et al. 2020]. We show how these algorithms can be leveraged for human capture, to simultaneously target high precision at low latency and memory cost, through a careful sparse coarse-to-fine scheme and reconfiguration of the neural differential rendering chain, with parsimonious encoding of human scenes and BRDFs. Our algorithm in turn achieves unprecedented resolution, pixel resolvability and performance for multi-view human surface capture by all metrics: distance to 3D reference, input image distance, computation time, ray rendering time, and memory footprint, as measured against a representative set of state of the art surface capture algorithms.

Actually verifying real-world accuracy in the context of human surface acquisition proves to be a key challenge, as it requires a highly accurate geometric surface reference to compare against: typical real-world datasets provide it for generic objects [Aanæs et al. 2016], but it is seldom found for multi-view human datasets. We thus introduce the MVMannequin dataset alleviating these drawbacks, comprised of 14 dressed real-size human mannequins in casual pose, providing both sub-millimeter hand-scans of their geometry, and their input images obtained from 68 calibrated viewpoints. We specifically opt for dressed mannequins for their perfect stillness, avoiding any micro-motions such as breathing or muscle tension change, that would yield perturbations beyond this bound. We further demonstrate qualitative achievements of the algorithm on publicly available multi-view human capture sequences of the 4DHumanOutfit dataset [Armando et al. 2023].

- We introduce MVMannequin, a real-world dataset of 14 human-realistic dressed mannequins both male and female with varied clothing, providing both calibrated multi-view observations and a hand-scanned reference of sub-millimeter accuracy.
- We provide a scalable approach for human surface capture from multiple images, with a novel sparse dynamically updated hierarchical backbone, achieving 1.13mm average accuracy and sub-millimeter accuracy for 77% of surface sampled points on MVMannequin reference data.
- Code and data available at https://projects-morpheo.gitlabpages.inria.fr/MillimetricHumans/.

2 RELATED WORK

Neural surface reconstruction. Initial volumetric approaches [Lombardi et al. 2019; Mildenhall et al. 2021] have drawn attention to the powerful capabilities of neural radiance modeling and differential rendering, with recent iterations showing various strategies to make the representation magnitudes more efficient [Kerbl et al. 2023; Müller et al. 2022]. It is however non-trivial to extract surfaces from the resulting volumetric opacity fields, prone to holes and floaters due to lack of surface regularization. Some promising methods thus reparametrize Gaussian splats for alignment to an underlying surface representation [Guédon and Lepetit 2024; Huang et al. 2024; Yu et al. 2024]. Yet, state of the art surface reconstruction performance is achieved by a concurrent family of neural methods that parametrize the shape implicitly with a signed-distance function (SDF) [Wang et al. 2021; Yariv et al. 2021]. Various improvements to boost the accuracy of these methods exist, *e.g.* [Darmon et al. 2022; Fu et al. 2022] incorporate a patch-based photo-consistency loss inspired by classic multi-view stereo algorithms, though at a significantly increased computational cost. [Li et al. 2023b] achieves outstanding accuracy on large scale scenes thanks to a combination of hash grids and numerical gradients, but trades speed for detail. Closest to our objectives, [Wang et al. 2023] combine hierarchical hash grids and implicit surface parametrization to achieve both accurate reconstruction and fast training, and [Wu et al. 2023] achieve an impressive combination of high quality rendering and geometric accuracy using a dense voxel grid. All these approaches have yet to be thoroughly validated in the context of human surface capture.

Neural pipeline optimization. Various works are relevant to the scalability and efficiency of the neural surface pipeline. The use of regular grids [Fridovich-Keil et al. 2022; Sun et al. 2022; Wu et al. 2023], hierarchical hash tables [Müller et al. 2022; Wang et al. 2023] or lower-dimensional component factorization [Cao and Johnson 2023; Chen et al. 2022, 2023; Fridovich-Keil et al. 2023] were all designed to reduce the size of the MLP needed to predict color and opacity/SDF, by parametrizing it with a set of local features instead of using one scene-global network. Hierarchical hashing and factorization, later unified with the works of [Gao et al. 2023], add a level of sparsification of the neural parameters, while making approaches more compute-friendly and easier to implement. Accelerating the ray-querying of these structures has also been a topic of significant interest, as it usually benefits both training and rendering applications. Again, hierarchical grids [Yu et al. 2021] or hash tables [Müller et al. 2022] can be leveraged to this end, along with reparametrization of the angular query component with spherical harmonic projection [Yu et al. 2021] or with a small MLP [Reiser et al. 2021]. Rendering-specific optimizations also exist, *e.g.* baking a pre-learned scene into a sparse grid of features, and deferring the MLP color conversion to the pixel domain by performing ray integration on MLP features instead of color [Hedman et al. 2021]. Sparse hierarchical structures may benefit volumetric differential approaches [Liu et al. 2020], and yield efficient volumetric query implementations on GPU [Kim et al. 2024]. Our approach unifies these different training, render and representation optimizations, transposes them to neural SDF approaches, and introduces sparse hierarchies that can be updated on-demand at training time.

Human prior models for capture. For completeness, we briefly mention that the geometry estimation problem can be tackled with human prior shape models, *e.g.* by fitting articulated [Habermann

et al. 2020], parametric [Loper et al. 2015], neural parametric [Kwon et al. 2021; Palafox et al. 2021] or part-based models [Osman et al. 2022]. Such model-based approaches are valuable for many applications, *e.g.* to drive and constrain capture of appearance-realistic image-based avatars, where face, hair, or clothing are explicitly modeled as additional rendering layers [Saito et al. 2023; Zheng et al. 2023], or added geometric surface detail layers which can be physics-aware [Casado-Elvira et al. 2022; Li et al. 2021; Pons-Moll et al. 2017], or both [Habermann et al. 2023], under tight functional constraints *e.g.* real-time [Habermann et al. 2019] or monocular input [Wang et al. 2022]. Yet generally they pull estimation toward a higher level prior distribution of shapes and away from measurement of very high frequency detail by construction. We pursue a different direction where such accuracy is the primary objective, notwithstanding the interest in semantically richer shape models.

3D validation datasets. Validating the quality of multi-view 3D reconstructions was pioneered with the Middlebury Multi-View Stereo Benchmark [Seitz et al. 2006], which introduced the notions of *accuracy* and *completeness* as Chamfer distances to and from the reconstructed output *w.r.t* a laser-scanned reference. Follow-up datasets increase resolution and object variety [Aanæs et al. 2016], or deal with open scenes [Knapitsch et al. 2017; Schöps et al. 2017; Strecha et al. 2008; Yao et al. 2020], with the same validation methodology, providing both registered input views and reference scans of orthogonal modality. While very useful, these datasets do not match typical full body human acquisition setups, where wider fields of view are used to allow for movement, and humans are imaged onto a small projected sensor area. Concerning 3D human datasets, on one hand many provide parametric human model data [Fang et al. 2021; Mahmood et al. 2019] or 3D body surface data obtained through photogrammetric or laser scans [Bogo et al. 2017; Ionescu et al. 2014; Jinka et al. 2022; RenderPeople 2024], but do not provide matching captured camera images thereof to be used as input for reconstruction. Others do provide multi-view images and 3D reconstructions from these viewpoints, but lack the independently measured comparative 3D reference [Cheng et al. 2023; Habermann et al. 2021; Işık et al. 2023; Wang et al. 2024; Xiong et al. 2024; Zheng et al. 2022]. Finally, some datasets build on the 3D models of the above 3D human datasets as ground truth, providing synthetically rendered viewpoints [Ge et al. 2024; Varol et al. 2017; Yang et al. 2023] but are intrinsically limited for our purpose due to their domain and use case gaps. More importantly, most come with undesirable caveats built-in: *i.e.* their ground truth lacks millimetric geometric surface detail due to decimated resolution [RenderPeople 2024], or is itself reconstructed from multi-view photogrammetry with the same measurement bias as the algorithms we need to evaluate. While all mentioned datasets are useful to train human AI models thanks to the wide subject, clothing and pose variability they provide, they are thus less than ideal for our validation task. This motivates our MVMannequin dataset, providing both high density sub-millimetric 3D hand-scans of human-like bodies in varied clothing, and the corresponding input images for multi-view reconstruction.

3 METHOD

3.1 Preliminary Notions

Differential rendering models [Mildenhall et al. 2021; Wang et al. 2021] can be functionally described as follows. For a given *forward pass*, image pixel colors are predicted for a set of rays in four steps: (1) for each ray, originating from the camera center o through an image pixel in the viewing direction v expressed as $\{p(t) = o + t\,v \mid t \geq 0\}$, select a discrete set of M 3D scene point samples $\{p_i = o + t_i\,v | i = 1, ..., M, t_i < t_{i+1}\}$ on the ray according to a *ray sampling model*, (2) for each 3D point sample, fetch parameters from a *scene representation*, *e.g.* a feature grid, (3) decode them to a given 3D color and opacity, using *e.g.* shallow MLPs, (4) integrate them in ray marching order according to a *rendering model*. The latter is typically the volumetric blending model of NeRF and NeuS [Max 1995]: where the rendered color of a pixel is discretely approximated by the blended sum along its M sampled points on the ray:

$$\mathbf{C} = \sum_{i=1}^{M} \alpha_i T_i \mathbf{C}_i, \qquad \text{with } T_i = \prod_{j<i} (1 - \alpha_j) \tag{1}$$

where α_i is the opacity value, T_i is the accumulated transmittance. NeuS differs from NeRF in that α_i is decoded from an SDF field as a roughly Gaussian mass of width $1/s$ centered on its 0-level set:

$$\alpha_i = \max\left(\frac{\Phi_s(f(p_i)) - \Phi_s(f(p_{i+1}))}{\Phi_s(f(p_i))}, 0\right), \tag{2}$$

where f is the SDF function, $\Phi_s(x) = (1 + e^{-sx})^{-1}$ the Sigmoid function, with the s value learned or scheduled during training.

3D reconstructions can be obtained from measured images by inverting the differential rendering model under a set of regularizing losses (*e.g.* Eikonal) to obtain scene-specific parameters. An *estimation algorithm* iteratively applies the forward pass to a selected *batch of rays* and minimizes a *rendering loss* between their predicted pixel colors and measured values, by back-propagating the loss gradients to scene parameters using the chain rule (*backward pass*).

3.2 Key Technical Contributions

We design a NeuS-inspired neural differential approach to reconstruct human scenes, with all components geared toward drastically maximizing representable resolution, and compute, rendering and GPU memory efficiency. Our forward pass and losses are summarized in fig. 2. We additionally leverage background images devoid of the subject of interest, specifically available in this context.

Scene representation. Our approach proposes a novel fixed-depth, constant-overhead sparse voxel-tile hierarchy to store scene parameters (detailed in §3.3), leveraging the inherent sparsity of human motion scenes. Each leaf tile voxel contains a scalar corresponding to its signed distance to the underlying shape, and compactly allows to decode a view-dependent color from the outer product of three local feature planes. This provides an inherent memory and compute advantage over dense [Sun et al. 2022] or hashed grids [Wang et al. 2023] that store all voxels or scale resolutions on GPU, in turn allowing reconstructions of unprecedented resolutions.

Coarse-to-fine estimation. To avoid resolution limits inherent to fixed grid representations, some approaches [Sun et al. 2022; Wu

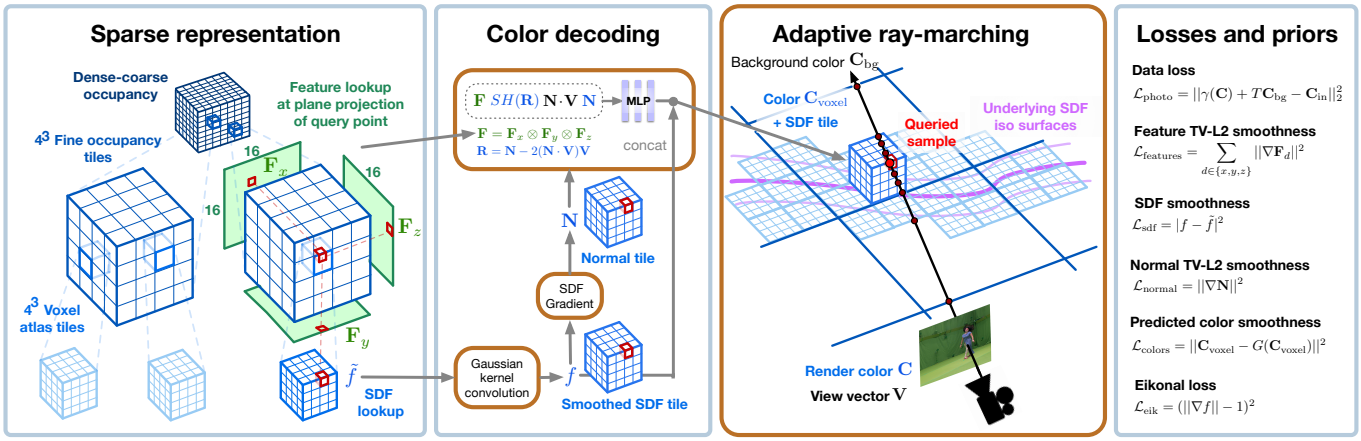

Fig. 2. Learnable features and SDF are stored as a sparse hierarchy, with factor decomposition of features at the intermediate level. All leaf voxel tiles are processed for inner SDF smoothness and normal computation. Color tiles are predicted using a tiny MLP with feature product, normal, normal and viewing ray vector dot product, and spherical harmonic projection of the reflected view vector, to preserve its high frequencies. An efficient ray marcher converts the SDF to opacities [Wang et al. 2021] and integrates the volume to produce pixel-wise color predictions. All orange boxes correspond to one GPU compute kernel in the forward and backward passes. The figure centers on a sample point query on a ray, and materializes corresponding features and voxels in red.

et al. 2023] refine by subsequently iterating over two grid resolutions. Similar to [Liu et al. 2020], we lift this restriction with arbitrary coarse-to-fine refinement (fig. 3), by periodically restarting with an upscaled resolution until a desired leaf resolution is reached. This leverages our fixed-overhead storage throughout, with memory scaling linearly in the number of occupied leaf tiles.

Deferring SDF decoding. to the ray-marching stage (§3.4) and *pre-computing decoded colors* per-voxel tile (§3.5) leads to a major remapping of forward pass stages on GPU with only 4 total kernel calls (fig. 2), and unlocks a set of desirable properties eluding previous approaches, such as arbitrary size ray-batching for full view sets, per-ray adaptive sampling, with single-kernel MLP-based color decoding and ray-marching. Pseudo-code of CPU/GPU calls is available in the supplementary.

3.3 Sparse Scene Representation

Flexibility is key at optimization time to closely adapt to an evolving shape estimate, while retaining sparsity. At rendering time, it is paramount to efficiently skip empty space to sample the volume parsimoniously. The requirements of sparsity and adaptability are challenging to fulfill simultaneously, with a range of possible representations analyzed in detail in the supplementary.

Sparse storage. We opt for an N^3 tree [Lefebvre et al. 2005; Museth 2013] as a sparse backbone structure, with the key originality that our proposal is both entirely managed on GPU and supports dynamic allocation. These trees generalize octrees in that each node has zero or N^3 children, with N the custom branching factor at each level. We empirically settle for a custom root node with larger N, akin to a coarse but dense grid, and two additional tree levels with $N = 4$, as shown in fig. 2. This configuration allows for better cache coherency with a constant number of indirections and only a small memory penalty compared to an octree. Each level can be packed in a 3D texture for hardware-accelerated lookups, with voxels grouped into leaf node tiles of size 4^3. These can contain decoded colors and the SDF as a single 4-tuple per voxel. The first two levels only encode occupancy at a coarse and finer level, and essentially store a grid of 3D indices pointing to a 4^3 tile into the next level.

Color feature reduction. We take inspiration from [Fridovich-Keil et al. 2023; Gao et al. 2023], and generate per-voxel color features by taking the outer product of three axis-aligned planes of features: $\mathbf{F} = \mathbf{F}_x \otimes \mathbf{F}_y \otimes \mathbf{F}_z$. Note that $\mathbf{F}$ can be computed on demand and doesn't require specific storage. This empirically reduces leaf-tile storage by 50% to 75% compared to naive per-voxel storage of $\mathbf{F}$.

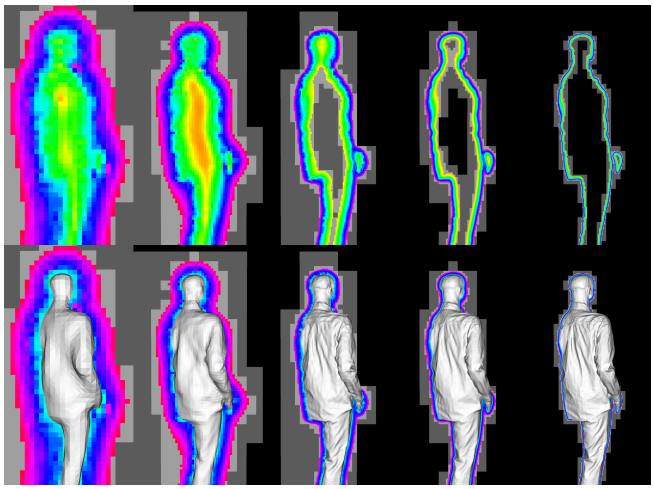

Fig. 3. Our method continuously prunes empty space as optimization progresses. Coarse occupancy tiles are shown in dark gray and fine occupancy tiles are in light gray. Signed-distance values are mapped to a hue to emphasize the level set regularization. The mesh extracted from the level set is shown in the bottom row [Lorensen and Cline 1987].

Dynamic allocations. We maintain a stack that keeps track of unused tiles in order to manage the memory pool from GPU-side operations only, with recursive updates to maintain a coherent hierarchy. This allows dynamic allocation and deletion of voxels or tiles anywhere in the volume after every optimization step, as illustrated fig. 3. We delete or introduce voxels based on conservative hysteresis thresholds of their SDF value, such that $|f| > \tau_{\max}$ are deleted whereas voxels with $|f| < \tau_{\min}$ trigger allocation of missing direct neighbors. We derive explicit values for $\tau_{\max}$ and $\tau_{\min}$ in the supplementary so that rays are guaranteed not to exit too early.

3.4 Adaptive Ray-Marching with Deferred SDF Decoding

Virtually all SOTA implementations including [Müller et al. 2022; Wu et al. 2023] bake scene lookups into tensors of dimension proportional to $R \times M \times F$, with R, M, F the number of batch rays, per-ray samples and features respectively. They are handily obtained from vectorized storage lookups and MLP queries on the full sample set of all batch rays. However, all intermediate operations, *e.g.* per MLP layer, generate GPU kernel calls and persistent tensors of similar size needed for backpropagation. This creates a large GPU memory burden that imposes practical limits on batch sizes (often 4096 rays).

In contrast, our approach avoids such overhead by pre-computing MLPs for every non-empty voxel tile beforehand in a temporary tile set, under a revised color decoding model discussed in §3.5. Ray-marching can then query colors and SDFs on-demand from temporary tiles, and perform late decoding of the opacity per eq. (2) with a single GPU kernel call over all rays and no ray-temporary tensors. Backpropagation can be performed similarly by accumulating gradients in a single set of voxel-tile temporaries instead of multiple memory intensive ray-batch tensors. We can thus process all rays of a given viewpoint at once, and keep iteratively accumulating gradients for arbitrary view sets in the same temporary tile set. To avoid naive uniform sampling, most of which would query empty space, our ray-marching dynamically skips empty tiles in our representation with a hierarchical version of the DDA algorithm [Museth 2014]. In non-empty regions, it can adapt integration step sizes on-the-fly with little overhead thanks to on-demand SDF access, instead of relying on expensive stratified sampling schemes or proposal networks, reviewed in [Li et al. 2023a].

3.5 Pre-Computed Color Decoding

Grid-based implementations [Wu et al. 2023] of the classic rendering model summarized in §3.1 typically compute each sample's color by querying an MLP with a trilinearly interpolated feature vector:

$$\mathbf{C}_{\text{classic}} = \mathit{render}(\mathit{MLP}(\mathit{lerp}(\mathit{grid}))). \tag{3}$$

This allows off-grid queries to produce spatially-varying colors at scales smaller than a voxel [Sun et al. 2022], but implies that MLP queries are done per sample point during the ray-integration stage. [Hedman et al. 2021] propose a deferred NeRF approach, hinting that the chain of operations can be re-ordered for other purposes:

$$\mathbf{C}_{\text{deferred}} = \mathit{MLP}(\mathit{render}(\mathit{lerp}(\mathit{grid})). \tag{4}$$

Here, the MLP is queried only once per ray over the volume integral of the interpolated grid features, to decode a pixel color. This strategy compresses and accelerates the rendering of already trained NeRF representations, but the latter is what we intend to improve. Instead, we propose to query the MLP once per voxel for a given viewpoint, and render with the result, interpolated at ray-sample locations:

$$\mathbf{C}_{\text{pre-computed}} = \mathit{render}(\mathit{lerp}(\mathit{MLP}(\mathit{grid}))). \tag{5}$$

First, this reordering unlocks the ray-marching simplification of §3.4. Second, the MLP can now solely focus on the angular distribution of the colors while their spatial distribution becomes fully encoded into the grid, better decoupling the two responsibilities. Consequently, given the restrained BRDF space of human scenes, we can reduce the size of the MLP and its computational cost as long as the voxels are small enough to explain the spatial distribution of the colors, which our sparse representation is designed to achieve (§3.3). We empirically demonstrate the validity of this intuition by evaluating our approach, using a 32-neuron MLP, against [Wu et al. 2023] and [Wang et al. 2023] that rely on the *classic* pipeline, in §4. More specifically, we parametrize the color prediction MLP as follows:

$$\mathbf{C}_{\text{voxel}} = \mathit{MLP}(\mathit{concat}(\mathbf{F}, \mathbf{N}, \mathbf{N} \cdot \mathbf{V}, \mathit{SH}(\mathbf{R}))) \tag{6}$$

with $\mathbf{F} = \mathbf{F}_x \otimes \mathbf{F}_y \otimes \mathbf{F}_z$ the color feature decoded from our sparse structure, $\mathbf{N} \triangleq \nabla f / ||\nabla f||$ the voxel's normal vector obtained with finite differences, $\mathbf{V}$ the view vector, and $\mathit{SH}(\mathbf{R})$ a frequency encoding of the reflected vector in the form of a spherical harmonic projection. Decomposing $\mathbf{R}$ is more important than other MLP input directions here, as it coincides with specular reflection directions. The entire color prediction is fused in a single GPU kernel to reduce data transfers and memory usage incurred by separate computations (see fig. 2). The only temporary storage needed is for the predicted color itself. An equivalent kernel implements the backward pass.

3.6 Regularizations and Losses

We detail the losses used for optimization in SDF, feature and color space, to constrain smoothness and explain the multi-view inputs.

Feature smoothness. is enforced directly on the feature planes with a TV-L2 loss. This promotes consistency between the neighboring features of different tiles : $\mathcal{L}_{\text{features}} = \sum_{d \in \{x,y,z\}} ||\nabla \mathbf{F}_d||^2$.

Spatial smoothness. Similarly to [Wu et al. 2023], we convolve the raw SDF $\tilde{f}$ with a 5^3 Gaussian kernel G to obtain a smoother SDF value $f = G(\tilde{f})$. Note that f is used for all subsequent operations. A regularization keeps f and $\tilde{f}$ close to each other: $\mathcal{L}_{\text{sdf}} = |f - \tilde{f}|^2$. The signed-distance property is promoted at each voxel by the Eikonal regularization $\mathcal{L}_{\text{eik}} = (||\nabla f|| - 1)^2$. The normal is also smoothed with $\mathcal{L}_{\text{normal}} = ||\nabla \mathbf{N}||^2$. Finally, we add a regularization on the predicted voxel colors $\mathcal{L}_{\text{colors}} = ||\mathbf{C}_{\text{voxel}} - G(\mathbf{C}_{\text{voxel}})||^2$, to avoid over-fitting to camera sensor noise.

Color correction. A significant practical issue arises with inter-camera exposure variations, which are difficult to globally control through camera hardware. We thus use a per-camera and per-channel color correction curve γ applied on the rendered colors $\mathbf{C}$. We parametrize $\gamma \colon [0,1] \longrightarrow [0,1]$ as a piece-wise linear function that we optimize along the rest of the parameters. We fix the endpoints to be 0 and 1 and apply a smoothness regularization to ensure that γ stays monotonous.

Rendering loss. We use a standard photometric loss, additionally accounting for per-camera color correction and background images if available. Since binary masks are never pixel accurate due to labeling error and intrinsic intra-pixel blending at occlusion boundaries, we blend in the background color $\mathbf{C}_{\text{bg}}$ as weighed by the residual ray transmittance, allowing for accurate sub-pixel performance:

$$\mathbf{C}_{\text{predicted}} = \gamma(\mathbf{C}) + T\mathbf{C}_{\text{bg}} \tag{7}$$

$$\mathcal{L}_{\text{photo}} = ||\mathbf{C}_{\text{predicted}} - \mathbf{C}_{\text{in}}||_2^2, \tag{8}$$

with $\mathbf{C}_{\text{in}}$ the observed color, $\mathbf{C}$ the rendered color without background. Following [Mildenhall et al. 2022], we apply a per channel weight on the photometric gradient $\left(\min\left(\mathbf{C}_{\text{predicted}}, \mathbf{C}_{\text{in}}\right) + 0.005\right)^{-1}$ so that bright and dark colors are treated in the same way.

4 EXPERIMENTS

We present the MVMannequin dataset, designed to characterize and compare metric reconstruction performance in human scenes. It includes 14 outfits on full-size mannequins, 6 male and 8 female, each requiring one hour for paired acquisitions using the [Kinovis 2024] platform, as described below. We selected a diverse range of clothing outfits, featuring casual or formal clothing and shoes; long dresses, skirts, shorts, and pants; short-sleeved and long-sleeved tops, including jackets; and plain color versus textured (fig. 4).

4.1 Dataset Acquisition Protocol

To capture submillimeter accurate reference data for each of the 14 outfits we provide, we first perform the hand scanning then the multi-camera shot back to back, to avoid any unwanted pose discrepancies. After this, we acquire background images and proceed with geometric calibrations of the camera set, which require shooting longer sequences of a calibration wand.

Scanning. We scan each instance using an Einscan HX hand scanner[1], in structured light mode with a manufacturer-specified 1mm resolution and 0.05mm accuracy. The scan was recorded in one take without need for sub-part repositioning. No filtering nor hole-filling was used for acquired points clouds. The resulting meshes of all mannequins are shown in fig. 4 and supplemental video.

Multi-camera acquisition. We capture sequences of 10 frames for each outfit using 68 calibrated RGB cameras of 4 megapixel resolution, with focal lengths between 8 and 16mm. The cameras are positioned roughly on a half-ellipsoid with radii 4m and 5m and height 5m looking towards the stage center (fig. 5), for an average image resolution of 2.5mm per pixel at the scene center. The total capture surface covers a length of 5.5m and a width of 3.5m, representative of a significant human movement area.

4.2 Baselines

We evaluate the quality of our reconstructions against a set of state-of-the-art baselines, two neural NeuS2 [Wang et al. 2023] and Voxurf [Wu et al. 2023], and two established multi-view stereo implementations, Colmap [Schönberger and Frahm 2016] and [RealityCapture 2024] (RC). We also evaluate against three splatting-based methods,

[1]https://www.einscan.com/einscan-hx/

Fig. 4. Hand-scanned mannequins. Female (top), male (bottom).

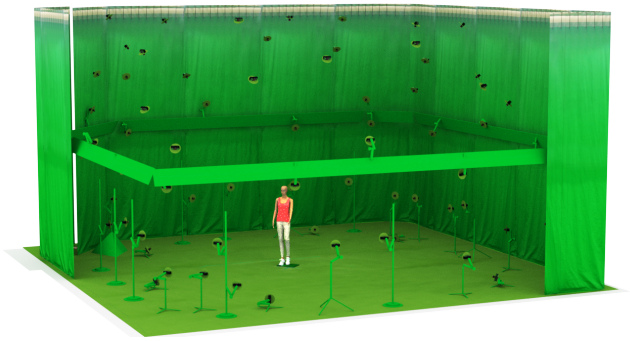

Fig. 5. The Kinovis multi-camera platform used to capture the dataset.

the original 3D Gaussian splatting (3DGS) [Kerbl et al. 2023], and its recent surface-capable variants 2D Gaussian splatting (2DGS) [Huang et al. 2024] and Gaussian opacity fields (GOF) [Yu et al. 2024]. A summary of each baseline's characteristics is given in table 1.

Dataset preparation. First, we temporally average then undistort the images and the backgrounds. We compute masked images using the backgrounds, as input to all baselines that require them (table 1, first row). All baselines use the same calibration except for RealityCapture since there was no simple way to provide one. The

Table 1. Indicative baselines characteristics

Baselines	Ours	NeuS2	Voxurf	RC	Colmap	3DGS	2DGS	GOF
Masks	both	✓	✓	✗	✗	✗	✗	✗
Full scene	✗	✗	✗	✓	✓	✓	✓	✓
Train	2min	5min	15min	5min	>1h	45min	1h20	3h30
VRAM	5GB	5GB	30GB	-	-	30GB	45GB	20GB
Framerate	~250	~15	<1	-	-	>300	>300	>300

Table 2. Completeness (mm), Accuracy (mm) and PSNR (db) over all scans. See supplemental document for a breakdown.

Comparisons									Ablations					
Metrics	Ours	Voxurf	NeuS2	RealityCapture	Colmap	3DGS	2DGS	GOF	r/2	r/4	r/8	w/ masks	w/o color correct.	w/o color smoothing
Completeness (mm)	**1.16**	1.50	2.51	4.90	3.67	-	2.29	3.27	1.63	1.84	6.97	1.68	2.56	1.20
Accuracy (mm)	**1.13**	1.68	1.97	3.64	3.36	-	4.41	2.81	1.44	1.75	5.11	1.54	2.32	1.14
PSNR (dB)	**36.33**	35.51	30.80	-	-	36.12	34.89	34.05	33.20	31.21	26.34	36.14	32.34	36.41

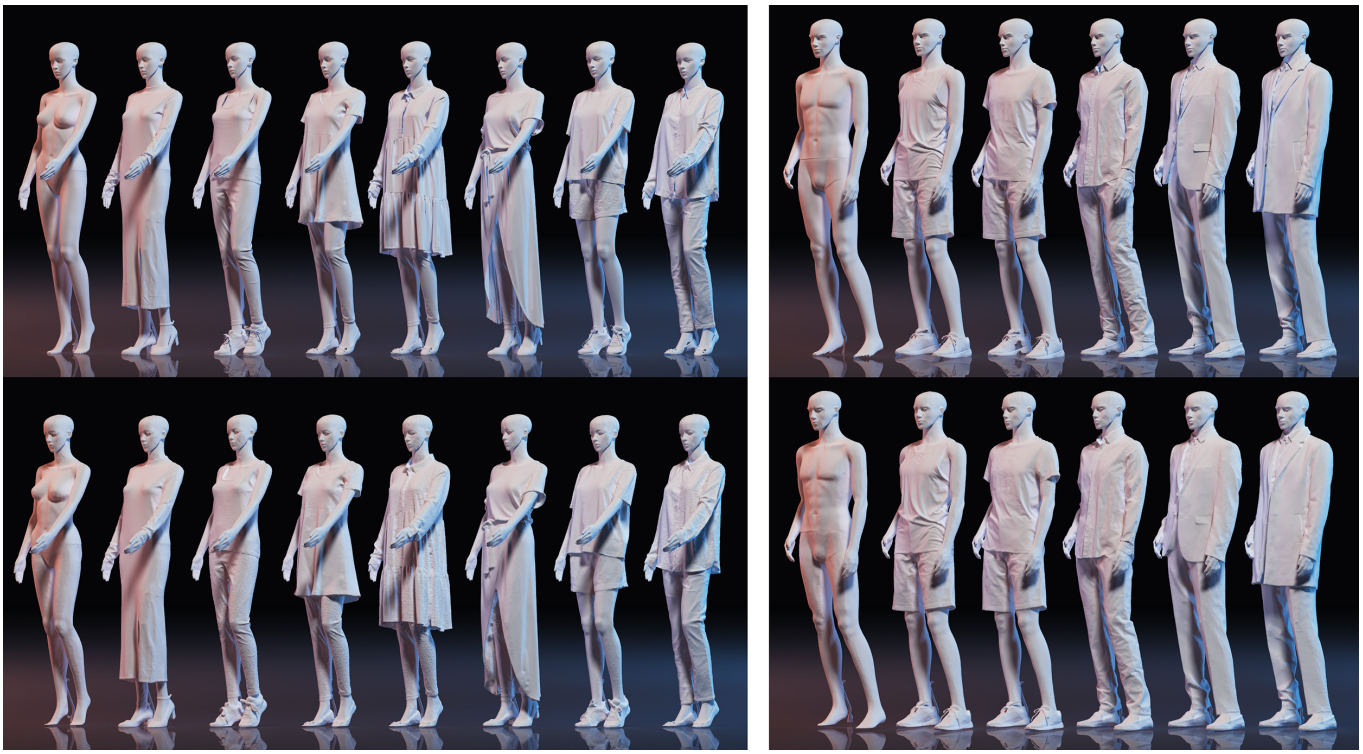

Fig. 6. Reference scans (top rows) versus our reconstructions (bottom rows) for our 8 female (left) and 6 male (right) mannequins dataset

sparse point cloud generated by Colmap is given to the splatting-based methods, as required. We disable Colmap's refining of camera intrinsics and extrinsics for fairness with the other baselines.

Geometric evaluation. We clip both the 3D reference and 3D reconstructions at a fixed height to remove the ground. A per-baseline robust ICP [Babin et al. 2019] accounts for the global registration bias before computing two-way point-to-mesh distances, less sensitive to mesh discretizations than point-to-point distances.

Photometric evaluation. The PSNR of the rendered images is computed within the reprojection of the scan, using the average ICP rigid transform of the best two baselines (Ours and Voxurf). The reprojection is eroded by two pixels since the silhouettes' borders are less accurate for the baselines relying on masks.

5 RESULTS

Quantitative results. We plot accuracy and completeness curves in fig. 7, showing the proportion of the vertices having a Euclidean distance to the reference that is less than $x$mm. We observe that our proposed method can reconstruct 77% of points with accuracy

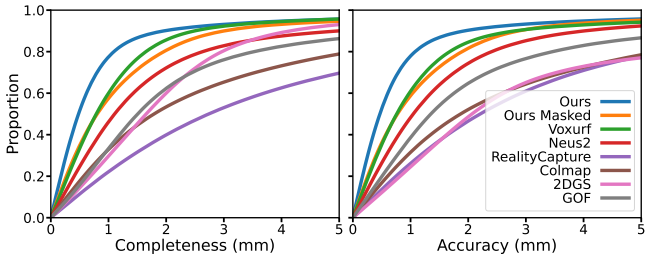

Fig. 7. Completeness and accuracy error curves

less than 1mm, while the second best method achieves about 60%, a 28% gain against SOTA. These results show the strong ability of our method to reconstruct geometric details with sub-millimeter accuracy and few outliers, with only 7% of points above 3mm distance from reference scan. RealityCapture and Colmap achieve an acceptable performance in well-textured areas, but tend to fail in uniform or specular regions of the surface, with less detail overall. Aggregate metrics are shown in table 2 (left). Both our accuracy

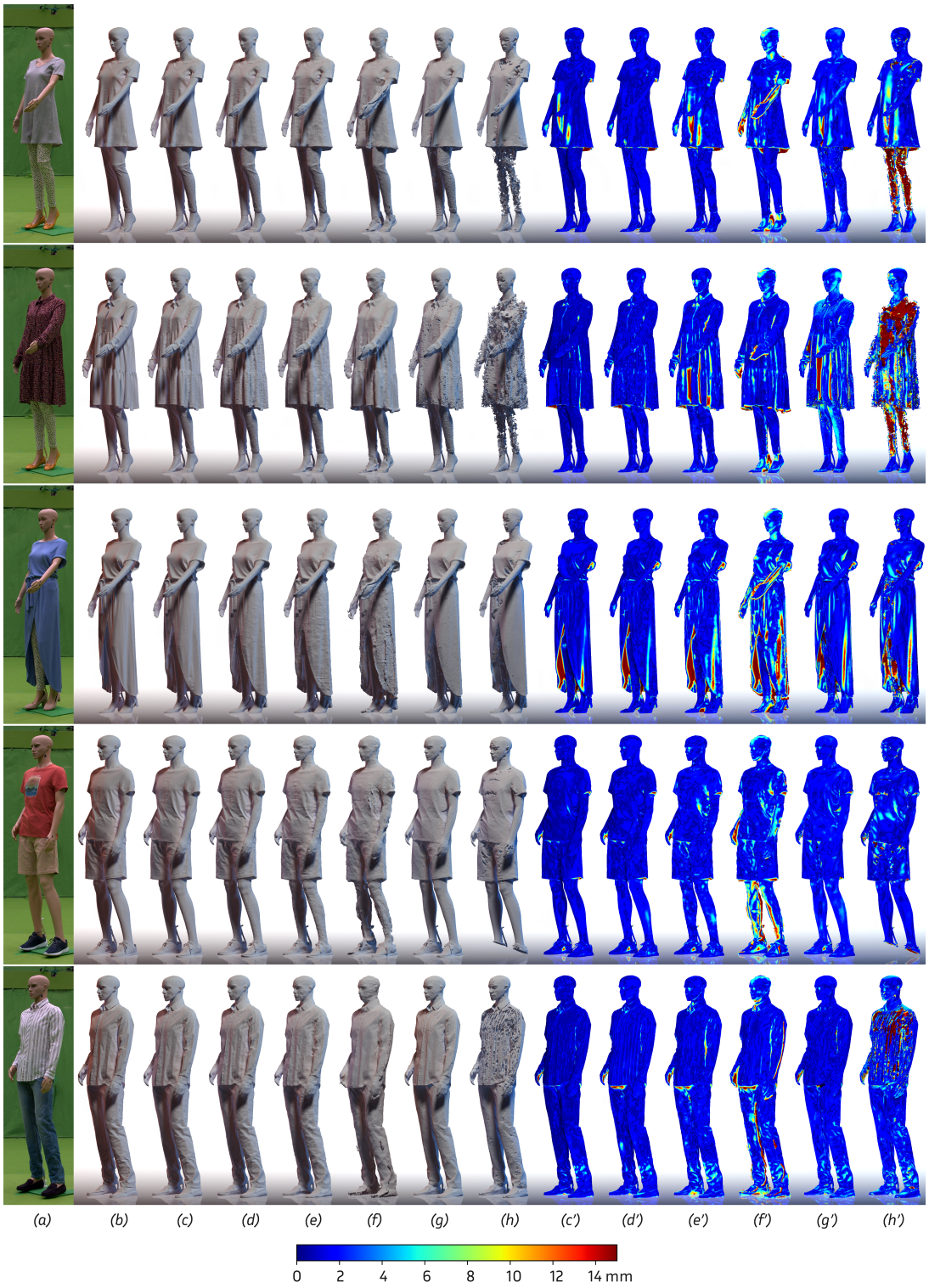

Fig. 8. Reconstructions on a subset of our proposed dataset with (a) input image, (b) reference scan, (c) our method, (d) Voxurf, (e) NeuS2, (f) Colmap, (g) 2DGS, (h) GOF and their associated accuracy heatmaps.

Fig. 9. Results of our reconstructions on some sequences of the 4DHumanOutfit dataset [Armando et al. 2023]

and completeness are below 1.2 mm with a PSNR slightly better than Voxurf's. 3DGS obtains a PSNR equivalent to ours thanks to its accurate specularities, but tends to aggregate many anisotropic Gaussians to explain high frequency texture patterns. This severely impacts the geometry of GOF and to a lesser extent 2DGS in those areas, because the Gaussians lack a cohesive orientation.

Error visualization. Fig. 8 provides a visual rendition of the error distribution on the surface, which is globally lower in our case, although Voxurf is able to better carve some of the concave regions. NeuS2 exhibits high frequency noise while Colmap is particularly sensitive to specularities on the arms, legs and head of the mannequins. GOF fails to extract a well-defined surface in the high texture frequency regions but achieves good accuracy otherwise, although noisy. We refer the reader to the supplemental document for a more detailed analysis as well as a detailed breakdown of the results. All our reconstructions are shown in fig. 6.

Ablations. We ablate the quality improvement per coarse-to-fine detail level, training with masked images instead of full background images, exposure correction, and color smoothness in table 2 (right). The color smoothness term avoids over-fitting to the sensor noise with a slightly decreased PSNR but a higher perceived quality (see supp. mat. for qualitative comparison). Color correction is important both for convergence of the geometry and rendering quality, explaining the inter-camera color variations. Training with background images improves metric performance by 30% over binary masks, as it allows modeling intra-pixel occlusion blending and avoids inherent inaccuracies of the masks. Our method is on par with Voxurf in terms of geometric quality with masked images, but with a higher PSNR, and outperforms Voxurf when background images are used. Each subsequent level of refinement (denoted as r/8, r/4, r/2 and Ours) is trained with images at twice the resolution and with voxels half the size, showing clear improvements in quality.

Qualitative results. We provide results on a representative set of motion sequences in clothing from the publicly available 4DHumanOutfit dataset [Armando et al. 2023]. The selected sequences exhibit running, boxing, crouching, jumping, dancing and a cartwheel motion, all covering several meters and varied poses. Fig. 9 and fig. 1 illustrate the reconstructions obtained from these sequences. Animated reconstructions are presented in the accompanying video. Note the abundance of surface detail on the faces, hair in motion, and clothing. Folds and creases are accurately reconstructed except in some concavities with uniform and ambiguous color. Notably, the micro-geometry of clothing seams is also captured in our reconstructions, as can be seen on various pants worn by the actors, and the noise levels on the surfaces are low enough for these details to be consistently represented frame after frame, despite the fact that our method does not apply any temporal smoothness, or estimate any temporal surface alignment.

6 CONCLUSION

We have shown that an efficient sparse structure combined with a streamlined optimization scheme can achieve state-of-the-art results in terms of geometric and photometric quality with significant memory and compute parsimony. Training only takes a few minutes while the scene can be volumetrically rendered at several hundreds of frames per second out of the box, with a trained model size of about 50MB. More importantly, our approach demonstrates that

pushing the raw resolution limits of differential rendering, and focusing the neural component on minimal BRDF requirements, translates into unprecedented pixel resolving power and reconstructed surface quality. We identify two limitations of our approach: first, we train each detail level for a predetermined number of iterations whereas automatically detecting convergence could provide speedups in several instances. Second, our small MLP can be insufficient to reconstruct strong and varied appearance specularities, which leads to geometry artifacts such as planar polygonalizations on the mannequin arms, and may hamper performance on more general scenes. Nevertheless, our experiments also highlight the consistent surface quality of results from our method on representative human motion scenes. Our proposed approach thus opens the path to several new research directions, and underlines that efficient yet expressive representations of the angular color distribution, and further sparse compression of the structure, are of future interest. Our low memory consumption is also promising for 3D modeling of unbounded scenes. In addition, fast training and rendering paves the way for sparse 4D reconstruction of temporal sequences at very high resolutions.

ACKNOWLEDGMENTS

This work was supported by French government funding managed by the National Research Agency under the Investments for the Future program (PIA) grant ANR-21-ESRE-0030 (CONTINUUM). This work was partly supported by JSPS KAKENHI Grant Number JP23H03439. We thank Julien Pansiot for his contribution to the acquisition process and Kinovis platform management.

REFERENCES

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. 2016. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision* 120, 2 (April 2016), 153–168. https://doi.org/10.1007/s11263-016-0902-9

Matthieu Armando, Laurence Boissieux, Edmond Boyer, Jean-Sébastien Franco, Martin Humenberger, Christophe Legras, Vincent Leroy, Mathieu Marsot, Julien Pansiot, Sergi Pujades, Rim Rekik, Grégory Rogez, Anilkumar Swamy, and Stefanie Wuhrer. 2023. 4DHumanOutfit: A multi-subject 4D dataset of human motion sequences in varying outfits exhibiting large displacements. *Computer Vision and Image Understanding* 237 (2023), 103836. https://doi.org/10.1016/j.cviu.2023.103836

Philippe Babin, Philippe Giguère, and François Pomerleau. 2019. Analysis of Robust Functions for Registration Algorithms. 1451–1457. https://doi.org/10.1109/ICRA.2019.8793791

Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2017. Dynamic FAUST: Registering Human Bodies in Motion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2017.591

Ang Cao and Justin Johnson. 2023. HexPlane: A Fast Representation for Dynamic Scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 130–141. https://doi.org/10.1109/cvpr52729.2023.00021

Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. 2022. PERGAMO: Personalized 3D Garments from Monocular Video. *Computer Graphics Forum* 41, 8 (Dec. 2022), 293–304. https://doi.org/10.1111/cgf.14644

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. In *European Conference on Computer Vision*. Springer, Springer Nature Switzerland, 333–350. https://doi.org/10.1007/978-3-031-19824-3_20

Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. 2023. Dictionary Fields: Learning a Neural Basis Decomposition. *ACM Transactions on Graphics* 42, 4 (July 2023), 1–12. https://doi.org/10.1145/3592135

Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. 2023. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv51070.2023.01829

François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. 2022. Improving neural implicit surfaces geometry with patch warping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52688.2022.00616

Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. 2021. Reconstructing 3D Human Pose by Watching Humans in the Mirror. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr46437.2021.01262

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 12479–12488. https://doi.org/10.1109/cvpr52729.2023.01201

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5501–5510. https://doi.org/10.1109/cvpr52688.2022.00542

Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/16415eed5a0a121bfce79924db05d3fe-Abstract-Conference.html

Yasutaka Furukawa and Jean Ponce. 2010. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 8 (Aug. 2010), 1362–1376. https://doi.org/10.1109/TPAMI.2009.161

Quankai Gao, Qiangeng Xu, Hao Su, Ulrich Neumann, and Zexiang Xu. 2023. Strivec: Sparse Tri-Vector Radiance Fields. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 17569–17579. https://doi.org/10.1109/iccv51070.2023.01611

Yongtao Ge, Wenjia Wang, Yongfan Chen, Hao Chen, and Chunhua Shen. 2024. 3D Human Reconstruction in the Wild with Synthetic Data Using Generative Models. *CoRR* abs/2403.11111 (2024). https://doi.org/10.48550/ARXIV.2403.11111 arXiv:2403.11111

Antoine Guédon and Vincent Lepetit. 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5354–5363.

Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhöfer, and Christian Theobalt. 2023. HDHumans: A Hybrid Approach for High-fidelity Digital Humans. *Proc. ACM Comput. Graph. Interact. Tech.* 6, 3 (2023), 36:1–36:23. https://doi.org/10.1145/3606927

Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021. Real-time Deep Dynamic Characters. *ACM Transactions on Graphics* 40, 4, Article 94 (Aug. 2021), 16 pages. https://doi.org/10.1145/3476576.3476653

Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. 2019. LiveCap: Real-Time Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 38, 2 (2019), 14:1–14:17. https://doi.org/10.1145/3311970

Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. 2020. DeepCap: Monocular Human Performance Capture Using Weak Supervision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 5051–5062. https://doi.org/10.1109/CVPR42600.2020.00510

Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. 2021. Baking Neural Radiance Fields for Real-Time View Synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv48922.2021.00582

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, Andres Burbano, Denis Zorin, and Wojciech Jarosz (Eds.). ACM, 32. https://doi.org/10.1145/3641519.3657428

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics* 42, 4 (July 2023), 1–12. https://doi.org/10.1145/3592415

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (July 2014), 1325–1339. https://doi.org/10.1109/tpami.2013.248

Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and P. J. Narayanan. 2022. SHARP: Shape-Aware Reconstruction of People in Loose Clothing. *International Journal of Computer Vision* 131, 4 (Dec. 2022), 918–937. https://doi.org/10.1007/s11263-022-01736-z

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions*

on Graphics 42, 4 (2023), 139:1–139:14. https://doi.org/10.1145/3592433

Doyub Kim, Minjae Lee, and Ken Museth. 2024. NeuralVDB: High-resolution Sparse Volume Representation using Hierarchical Neural Networks. *ACM Transactions on Graphics* 43, 2, Article 20 (Feb. 2024), 21 pages. https://doi.org/10.1145/3641817

Kinovis. 2024. Kinovis, Inria 4D Modeling Multi-Camera Platform. Online https://kinovis.inria.fr/.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–13. https://doi.org/10.1145/3072959.3073599

Youngjoon Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2021. Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 24741–24752. https://api.semanticscholar.org/CorpusID:237513692

Sylvain Lefebvre, Samuel Hornus, and Fabrice Neyret. 2005. Texture sprites: texture elements splatted on surfaces *(ACM Conferences)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1053427.1053454 Title from The ACM Digital Library.

Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. 2023a. NerfAcc: Efficient Sampling Accelerates NeRFs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv51070.2023.01699

Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. 2021. Deep Physics-aware Inference of Cloth Deformation for Monocular Human Performance Capture. In *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*. IEEE, 373–384. https://doi.org/10.1109/3DV53792.2021.00047

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023b. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52729.2023.00817

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural Sparse Voxel Fields. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/b4b758962f17808746e9bb832a6fa4b8-Abstract.html

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Transactions on Graphics* 38, 4, Article 65 (July 2019), 14 pages. https://doi.org/10.1145/3306346.3323020

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

William E. Lorensen and Harvey E. Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm.. In *SIGGRAPH*, Maureen C. Stone (Ed.). ACM, 163–169. https://doi.org/10.1145/280811.281026

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 5442–5451. https://doi.org/10.1109/iccv.2019.00554

N. Max. 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* 1, 2 (June 1995), 99–108. https://doi.org/10.1109/2945.468400

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. 2022. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52688.2022.01571

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (Dec. 2021), 99–106. https://doi.org/10.1145/3503250

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics* 41, 4 (July 2022), 1–15. https://doi.org/10.1145/3528223.3530127

Ken Museth. 2013. VDB: High-resolution sparse volumes with dynamic topology. *ACM Transactions on Graphics* 32, 3 (June 2013), 1–22. https://doi.org/10.1145/2487228.2487235

Ken Museth. 2014. Hierarchical digital differential analyzer for efficient ray-marching in OpenVDB. In *ACM SIGGRAPH 2014 Talks (SIGGRAPH '14)*. ACM. https://doi.org/10.1145/2614106.2614136

Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2022. SUPR: A Sparse Unified Part-Based Human Body Model. In *European Conference on Computer Vision (ECCV)*. https://supr.is.tue.mpg.de

Pablo R. Palafox, Aljaz Bozic, Justus Thies, Matthias Nießner, and Angela Dai. 2021. NPMs: Neural Parametric Models for 3D Deformable Shapes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 12675–12685. https://doi.org/10.1109/ICCV48922.2021.01246

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. 2017. ClothCap: seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–15. https://doi.org/10.1145/3072959.3073711

RealityCapture. 2024. RealityCapture. https://www.capturingreality.com/ https://www.capturingreality.com/.

Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. 2021. KiloNeRF: Speeding up Neural Radiance Fields with Thousands of Tiny MLPs. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv48922.2021.01407

RenderPeople. 2024. RenderPeople. https://renderpeople.com/

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2023. Relightable Gaussian Codec Avatars. *CoRR* abs/2312.03704 (2023). https://doi.org/10.48550/ARXIV.2312.03704 arXiv:2312.03704

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2016.445

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2538–2547. https://doi.org/10.1109/cvpr.2017.272

S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. 519–528. https://doi.org/10.1109/CVPR.2006.19

C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8. https://doi.org/10.1109/cvpr.2008.4587706

Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022. Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5459–5469. https://doi.org/10.1109/cvpr52688.2022.00538

Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from Synthetic Humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2017.492

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 27171–27183. https://proceedings.neurips.cc/paper/2021/hash/e41e164f7485ec4a28741a2d0ea41c74-Abstract.html

Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. 2022. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII (Lecture Notes in Computer Science, Vol. 13692)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 1–19. https://doi.org/10.1007/978-3-031-19824-3_1

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 2024. 4D-DRESS: A 4D Dataset of Real-world Human Clothing with Semantic Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. 2023. NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 3295–3306. https://doi.org/10.1109/iccv51070.2023.00305

Tong Wu, Jiaqi Wang, Xingang Pan, Xudong Xu, Christian Theobalt, Ziwei Liu, and Dahua Lin. 2023. Voxurf: Voxel-based Efficient and Accurate Neural Surface Reconstruction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=DSy8tP4WctmZ

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, Shuguang Cui, and Xiaoguang Han. 2024. MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19801–19811.

Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei

Liu, and Lei Yang. 2023. SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 20282–20292. https://doi.org/10.1109/iccv51070.2023.01855

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. 2020. BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr42600.2020.00186

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume Rendering of Neural Implicit Surfaces. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 4805–4815. https://proceedings.neurips.cc/paper/2021/hash/25e2a30f44898b9f3e978b1786dcd85c-Abstract.html

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1a77befc3b608d6ed363567685f70e1e-Abstract.html

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 5752–5761. https://doi.org/10.1109/iccv48922.2021.00570

Zehao Yu, Torsten Sattler, and Andreas Geiger. 2024. Gaussian Opacity Fields: Efficient and Compact Surface Reconstruction in Unbounded Scenes. *CoRR* abs/2404.10772 (2024). https://doi.org/10.48550/ARXIV.2404.10772 arXiv:2404.10772

Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. 2022. Structured Local Radiance Fields for Human Avatar Modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52688.2022.01543

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023. AvatarReX: Real-time Expressive Full-body Avatars. *ACM Transactions on Graphics* 42, 4 (July 2023), 1–19. https://doi.org/10.1145/3592101