HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture

Ziyan Wang[1,3] Giljoo Nam[3] Tuur Stuyck[3]
Stephen Lombardi[3] Michael Zollhöfer[3] Jessica Hodgins[1,2] Christoph Lassner[3]
[1]Carnegie Mellon University [2]Meta AI [3]Reality Labs Research

Abstract

Capturing and rendering life-like hair is particularly challenging due to its fine geometric structure, the complex physical interaction and its non-trivial visual appearance. Yet, hair is a critical component for believable avatars. In this paper, we address the aforementioned problems: 1) we use a novel, volumetric hair representation that is composed of thousands of primitives. Each primitive can be rendered efficiently, yet realistically, by building on the latest advances in neural rendering. 2) To have a reliable control signal, we present a novel way of tracking hair on the strand level. To keep the computational effort manageable, we use guide hairs and classic techniques to expand those into a dense hood of hair. 3) To better enforce temporal consistency and generalization ability of our model, we further optimize the 3D scene flow of our representation with multiview optical flow, using volumetric raymarching. Our method can not only create realistic renders of recorded multi-view sequences, but also create renderings for new hair configurations by providing new control signals. We compare our method with existing work on viewpoint synthesis and drivable animation and achieve state-of-the-art results. https://ziyanw1.github.io/hvh/

1. Introduction

Although notable progress has been made towards the realism of human avatars, cephalic hair is still one of the hardest parts of the human body to capture and render: with usually more than a hundred-thousand components, with complex physical interaction among them and with complex interaction with light, which is extraordinarily hard to model. However, it is an important part of our appearance and identity: hair styles can convey everything from religious beliefs to mood or activity. Hence, hair is critically important to make virtual avatars believable and universally usable.

Previous work on mesh based representations [3, 20, 25, 52, 57, 58, 68] has shown promising results on modeling the face and skin. However, they suffer when modeling hair, because meshes are not well suited for representing hair geometry. Recent volumetric representations [26, 35] have high DoF which allows modeling of a changing geometric structure. They have achieved impressive results in 3D scene acquisition and rendering from multi-view photometric information. Compared to other geometric representations like multi-plane images [2, 5, 34, 56, 76] or point-based representations [1, 19, 33, 49, 65], volumetric representations support a larger range of camera motion for view extrapolation and do not suffer from holes when rendering dynamic geometry like point-based representations. Furthermore, they can be learned from multi-view RGB data using differentiable volumetric ray marching, without additional MVS methods.

However, one major flaw of volumetric representations is their cubic memory complexity. This problem is particularly significant for hair, where high resolution is a requirement. NeRF [35] circumvents the $O(n^3)$ memory complexity problem by parameterizing a volumetric radiance field using an MLP. Given the implicit form, the MLP-based implicit function is not limited by spatial resolution. A hierarchical structure with a coarse and fine level radiance function is used and an importance resampling based on the coarse level radiance field is utilized for boosting sample resolution. Although promising empirical results have been shown, they come with at the advance of high rendering time and the quality is still limited by the coarse level sampling resolution. Another limitation of NeRFs is that they were initially designed for static scenes. There is some recent work [21, 22, 41, 42, 46, 59, 63, 67, 74] that extends the original NeRF concept to modeling dynamic scenes. However, they are still limited to relatively small motions, do not support drivable animation or are not efficient for rendering.

We present a hybrid representation: by using many volumetric primitives, we focus the resolution of the model onto the relevant regions of the 3D space. For each of the volumes, we construct a neural representation that captures the local appearance of the hair in great detail, similar to [24, 27, 47, 63] . However, without explicitly modeling the dynamics and structure of hair, it would be hard for the model to learn these properties solely through the indirect supervision of the multi-view appearance. Given that the model learns to position primitives in an unsupervised manner, the model is also prone to overfitting as a result of not incorporating any temporal consistency during training. We address the problem of spatio-temporal modeling of dynamic upper head and hair by explicitly modeling hair dynamics at the coarse level and by enforcing temporal consistency of the model by multi-view optical flow at the fine level.

Procedurally, we first perform hair strand tracking at a coarse level by lifting multi-view optical flow to a 3D scene flow. To constrain the hair geometry and reduce the impact of the noise in multi-view optical flow, we also make sure the tracked hair strands preserve geometric properties like shape, length and curvature across time. As a second step, we attach volumes to hair strands to model the dynamic scene which can be optimized using differentiable volumetric raymarching. The volumes that are attached to the hair strands are regressed using a decoder that takes per-hair-strand features and a global latent code as input and is aware of the hair specific structure. Additionally, we further enforce fine 3D flow consistency by rendering the 3D scene flow of our model into 2D and compare it with the corresponding ground truth optical flow. This step is essential for making the model generalize better to unseen motions. To summarize, the contributions of this work are

- A hybrid neural volumetric representation that binds volumes to guide hair strands for hair performance capture.
- A hair tracking algorithm that utilizes multiview optical flow and per-frame hair strand reconstruction while preserving specific geometric properties like hair strand length and curvature.
- A volumetric ray marching algorithm on 3D scene flow which enables optimization of the position and orientation of each volumetric primitive through multiview 2D optical flow.
- A hair specific volumetric decoder for hair volume regression and with awareness of hair structure.

2. Related Work

In this section, we discuss the most closely related classical hair dynamic and shape modeling methods. We then discuss learning-based approaches that use either volumetric or non-volumetric scene representations for spatio-temporal modeling.

Image-based Hair Geometry Acquisition is challenging due to the complicated hair geometry, massive number of strands, severe self occlusion and collision and view-dependent appearance. Paris *et al.* [38, 39] and Wei *et al.* [64] reconstruct 3D hair geometry from 2D/3D orientation fields using multi-view images. Luo *et al.* [29, 31] further improve the 3D reconstruction by refining the point cloud from traditional MVS with structure-aware aggregation and strand-based refinement. Luo *et al.* [30] and Hu *et al.* [12] progressively fit hair specific structures like ribbons and wisps to the point cloud. Recently, Nam *et al.* [36] substitute the plane assumption in the conventional MVS by a line-based structure to reconstruct 3D line clouds. Sun *et al.* [55] use OLAT images for more efficient reconstruction of line-based MVS and develop an inverse rendering pipeline for hair that reasons about hair specific reflectance. However, none of those methods explicitly model temporal consistency for a time series capture.

Dynamic Hair Capture. Compared to the vast body of work on hair geometry acquisition, the work on hair dynamics [11, 69, 71, 75] acquisition is much less. Zhang *et al.* [75] uses hair simulation to enforce better temporal consistency over a per-frame hair reconstruction result. Hu *et al.* [11] solves the physics parameters of a hair dynamics model by running parallel processes under different simulation parameters and adopting the one that best matches the visual observation. Xu *et al.* [69] performs visual tracking by aligning per-frame reconstruction of hair strands with motion paths of hair strands on a horizontal slice of a video volume. Yang *et al.* [71] developed a deep learning framework for hair tracking using indirect supervision from 2D hair segmentation and a digital 3D hair dataset. However those methods mainly focus on geometry modeling and are not photometrically accurate or do not support drivable animation.

Non-Volumetric Representations are widely studied in the literature of spatio-temporal modeling Mesh-based representations [3, 20, 25, 57, 58, 68] are a perfect fit for modeling surfaces and highly efficient to render. However, they have limitations for modeling complex geometries like hair. Multi-plane images [2, 5, 34, 56, 76] are good at modeling continuous shapes similar to volumetric representations, but are limited to a constrained set of viewing angles. Point cloud representations [1, 19, 33, 49, 65] can model various geometries with high fidelity. When used for appearance modeling, however, point-based representations might suffer from their innate sparseness which might result in holes. Thus image-level rendering techniques [48] are often accompanied with such representations for completeness.

Volumetric Representations are highly flexible and thus can model many different objects. They are designed for geometric completeness given their dense grid-like structure. Many previous works have demonstrated the strength of such representations in geometry modeling [8, 9, 15, 60, 61, 66, 77]. Some recent works [26, 44, 53] have also shown their effectiveness in modeling appearance. DeepVoxels [53] learn a 3D grid of features as the scene representation. Neural volumes [26] learns a grid of discrete color and density values via volumetric raymarching. Neural body [44] incorporates SMPL [28] with Neural Volumes [26] for body modeling. Nevertheless, the rendering quality, efficiency and memory footprint of those volumetric representations is still limited by the voxel resolutions. To conquer this major drawback of volumetric methods, MVP [27] proposes a hybrid representation for efficient and high-fidelity rendering. It attaches a set of local volumetric primitives to a tracked head mesh and employs a tailored volumetric raymarching algorithm that is developed for fast rendering via a BVH [16]. The tracked mesh provides a good initialization for the positions and rotations of the primitives that are jointly learned. Still, finding the globally optimal positions and rotations purely based on a photometric reconstruction loss is highly challenging due to many local minima in the energy formulation.

Coordinate-based Representations have been the major focus of recent literature in 3D learning due to their low memory footprint and ability to dynamically assign the model capacity to the correct regions of 3D space. Many works have demonstrated their ability to reconstruct high fidelity geometry [6, 10, 14, 32, 40, 43, 50, 51] or to generate photo-realistic rendering results [24, 35, 37, 54, 72, 73]. NeRF [35] learns a volumetric radiance field of a static scene from multi-view photometric supervision using a differentiable raymarcher, but comes with a large rendering time. Several works [23, 24, 47, 73] have improved the rendering efficiency of NeRF on static scene. Among all those approaches, the most related to ours are spatio-temporal modeling techniques [21, 22, 41, 42, 46, 59, 63, 67]. Non-rigid NeRF [59], D-NeRF [46] and Nerfies [41] introduce a dynamic modeling framework with a canonical radiance field and per-frame warpings. Some works [21, 22, 62, 63, 67, 74] model a 3D video by additionally conditioning the radiance field on temporally varying latent codes or an additional time index. Xian *et al.* [67] further leverages depth as an extra source of supervision. STaR [74] models scenes that consist of a background and one dynamic rigid object. NSFF [22] also combines a static and dynamic NeRF pipeline and uses optical flow to constrain the 3D scene flow derived from the NeRF model of adjacent time frames. Wang *et al.* [63] introduce a grid of local animation codes for better generalization and improved rendering efficiency. However, these methods are still limited by either sampling resolution or ability to model complex motions and do not generalize well to unseen motions.

3. Method

In this section, we introduce our hybrid neural volumetric representation for hair performance capture. Our representation combines both, the drivability of guide hair strands and the completeness of volumetric primitives. Additionally, the guide hair strands serve as an efficient coarse level geometry for volumetric primitives to attach to, avoiding unnecessary computational expense on empty space. As a result of guide hair strand tracking as well as dense 3D scene flow refinement, our model is temporally consistent with better generalization over unseen motions. As illustrated in Fig. 1, the whole pipeline contains two major steps which we will explain separately. In the first step, we perform strand-level tracking that leverages multi-view optical flow information and propagates information about a subset of tracked hair strands into future frames. To save computation time, we track only guide hairs instead of tracking all hair strands. This is a widely used technique in hair animation and simulation [7, 13, 45], which leads to a significant boost in run time performance. However, getting the guide hairs tracked is not enough to model the hair motion and appearance or to animate all the hairs due to the sparseness of the guide hairs. To circumvent this, we combine it with a volumetric representation by attaching volumetric primitives to the nodes on the guide hairs. This hybrid representation has good localization of hairs in an explicit way and has full coverage of all the hairs, making use of the benefits of both representations. Another advantage is that the introduction of volumes allows optimizing hair shape and appearance by multi-view dense photometric information via differentiable volumetric ray marching. In the second step, we use the attached volumetric primitives to model the hairs that are surrounding the guide hair strands to achieve dense hair appearance, shape and motion acquisition. A hair specific volume decoder is designed for regressing those volumes, conditioning on both a global latent vector and hair strand feature vectors with hair structure awareness. Additionally, we develop a volumetric raymarching algorithm for 3D scene flow that facilitates the learning from multi-view 2D optical flow. We show in the experiments that the introduction of additional optical flow supervision yields better temporal consistency and generalization of the model.

3.1. Guide Hair Tracking

We frame the guide hair tracking process as an optimization problem. Given the guide hair strands and multi-view optical flow at the current frame t, we unproject and fuse optical flow under different camera poses into 3D flow and use that to infer the next possible position of the guide hairs at the next frame $t + 1$. The guide hair initialization at first frame is prepared by artist.

Data Setup and Notation. In our setting, we perform hair tracking using multi-view video data. We use a multi-camera system with around 100 synchronized color cameras that produces 2048×1334 resolution images at 30 Hz.

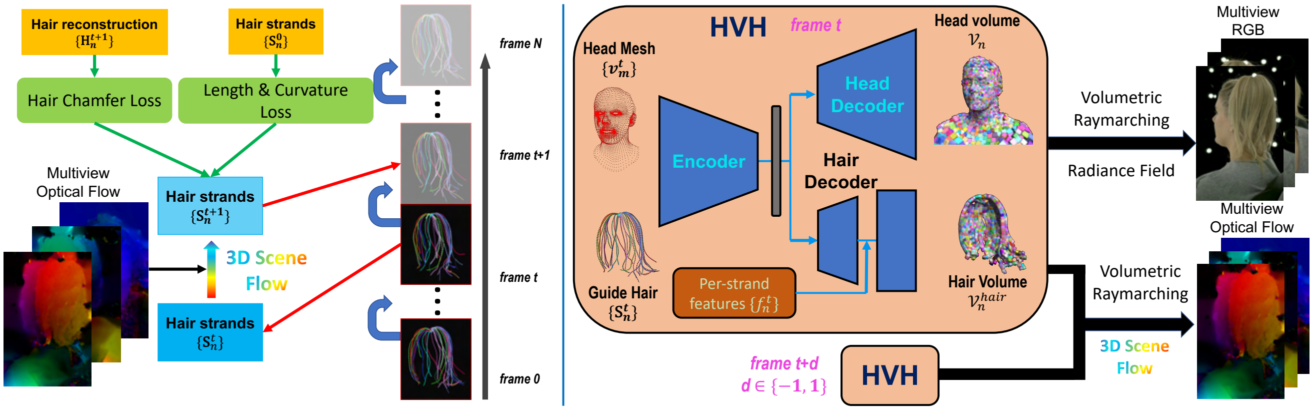

Figure 1. **Pipeline.** Our method consists of two stages: in the first stage, we perform guide hair tracking with multiview optical flow as well as per-frame hair reconstruction. In the second stage, we further amplify the sparse guide hair strands by attaching volumetric neural rendering primitives and optimizing them by using the multiview RGB and optical flow data.

The cameras are focused at the center of the capture system and distributed spherically at a distance of one meter to provide as many viewpoints as possible. Camera intrinsics and extrinsics are calibrated in an offline process. We generate multi-view optical flow between adjacent frames for each camera, using the OpenCV [4] implementation of [18]. We acquire per-frame hair geometry by running [36]. We parameterize guide hairs as connected point clouds. Given a specific hair strand $\mathbf{S}^t$ at time frame t, we denote the Euclidean coordinate of the $\mathbf{n}$th node on hair strand $\mathbf{S}^t$ as $\mathbf{S}_n^t$. Similarly, we have the future position of $\mathbf{S}_n^t$ at time frame $t+1$ as $\mathbf{S}_n^{t+1}$. Next we introduce the notations for multi-view camera related information. We denote $\Pi_i(\cdot)$ as the camera transformation matrix of camera i which projects a 3D point into 2D image coordinate. We denote $\mathbf{I}_{of,i}$ and $\mathbf{I}_{d,i}$ as 2D matrix of optical flow and depth of camera i respectively. We denote $\mathbf{H}_n^t$ as the reconstructed point cloud with direction from [36, 55]. Unless otherwise stated, all bold lower case symbols denote vectors.

Tracking Objectives. Given camera i, we could project a 3D point into 2D to retrieve its 2D image index. The camera projection is defined as

$$\hat{\mathbf{p}}_{s,i}^t = \begin{bmatrix} \mathbf{p}_{s,i}^t \\ \mathbf{1} \end{bmatrix} = \Pi_i(\mathbf{S}_n^t),$$

where $\hat{\mathbf{p}}_{s,i}^t$ is the homogeneous coordinate of $\mathbf{p}_{s,i}^t$. Given the camera projection formulation, we formulate the first data-term objective based on optical flow as follows:

$$\mathcal{L}_{of} = \sum_{n,i} \omega_{n,i} ||\mathbf{S}_n^{t+1} - \mathbf{Z}_i(\mathbf{S}_n^{t+1})\Pi_i^{-1}(\mathbf{p}_{s,i}^t + \delta_\mathbf{p})||_2^2,$$
$$\omega_{n,i} = exp(-\sigma||\mathbf{Z}_i(\mathbf{S}_n^t) - \mathbf{I}_{d,i}(\mathbf{p}_{s,i}^t)||_2^2),$$
$$\delta_\mathbf{p} = \mathbf{I}_{of,i}(\mathbf{p}_{s,i}^t),$$

where we denote $\mathbf{Z}_i(\cdot)$ as the function that represents the depth of a certain point under camera i and ω_i serves as a weighting factor for view selection where a smaller value means larger mismatch of projected depth and real depth under the ith camera pose. We use a $\sigma = 0.01$.

In parallel with the data-term objective on optical flow, we add another data-term objective to facilitate geometry preserved tracking, which compares the Chamfer distance between tracked guide hair strands and the per-frame hair reconstruction from [36]. This loss is designed to make sure that the guide hair geometry point cloud will not deviate too much from the true hair geometry. Unlike the conventional Chamfer loss, we also penalize the cosine distance between the directions of $\mathbf{S}_n^t$ and the direction of its closest $k = 10$ neighbors as $\mathcal{H}(\mathbf{S}_n^{t+1}) \subsetneq \{\mathbf{H}_n^{t+1}\}$; the losses are defined as:

$$\mathcal{L}_{hdir} = \sum_{n,\mathbf{h}\in\mathcal{H}(\mathbf{S}_n^{t+1})} \omega_{n,\mathbf{h}}^d (1 - |\cos(\mathbf{dir}(\mathbf{S}_n^{t+1}), \mathbf{dir}(\mathbf{h}))|),$$
$$\mathcal{L}_{hpos} = \sum_{n,\mathbf{h}\in\mathcal{H}(\mathbf{S}_n^{t+1})} \omega_{n,\mathbf{h}}^r ||\mathbf{S}_n^{t+1} - \mathbf{h}||_2^2,$$

where $\omega_{n,\mathbf{h}}^d = exp(-\sigma||\mathbf{S}_n^{t+1} - \mathbf{h}||_2^2)$ is a spatial weighting, $cos(\cdot,\cdot)$ is a cosine distance function between two vectors and $\mathbf{dir}(\mathbf{S}_n^{t+1}) = \mathbf{S}_{n+1}^{t+1} - \mathbf{S}_n^{t+1}$ is a first order approximation of the hair direction at $\mathbf{S}_n^{t+1}$. $\omega_{n,\mathbf{h}}^r = cos(\mathbf{dir}(\mathbf{S}_n^{t+1}), \mathbf{dir}(\mathbf{h}))$ is a weighting factor that aims at describing the direction similarity between $\mathbf{S}_n^{t+1}$ and $\mathbf{h}$. With $\mathcal{L}_{hdir}$, we could groom the guide hairs $\mathbf{S}_n^{t+1}$ to have similar direction to its closest $k = 10$ neighbors in $\mathcal{H}(\mathbf{S}_n^{t+1})$, resulting in a more consistent guide hair direction distribution. Alternatively, $\mathcal{L}_{hpos}$ guarantees that the tracked guide hairs do not deviate too much from the reconstructed hair shapes.

However, with just the data-term loss, the tracked guide hairs might overfit to noise in the data terms. To prevent this, we further introduce several model-term objectives for hair shape regularization.

$$\begin{aligned}\mathcal{L}_{len} &= \sum_n (||\mathbf{dir}(\mathbf{S}_n^{t+1})||_2 - ||\mathbf{dir}(\mathbf{S}_n^0)||_2)^2, \\ \mathcal{L}_{tang} &= \sum_n ((\mathbf{S}_{n+1}^{t+1} - \mathbf{S}_n^{t+1} - \mathbf{S}_{n+1}^t + \mathbf{S}_n^t) \cdot \mathbf{dir}(S_n^t))^2 + \\ &\quad ((\mathbf{S}_{n+1}^t - \mathbf{S}_n^t - \mathbf{S}_{n+1}^{t+1} + \mathbf{S}_n^{t+1}) \cdot \mathbf{dir}(S_n^{t+1}))^2, \\ \mathcal{L}_{cur} &= \sum_n (\mathbf{cur}(\mathbf{S}_n^{t+1}) - \mathbf{cur}(\mathbf{S}_n^0)),\end{aligned}$$

where $\mathbf{cur}(\mathbf{S}_n^t)$ is a numerical approximation of curvature at point $\mathbf{S}_n^t$ and is defined as:

$$\sqrt{\frac{24(||\mathbf{dir}(\mathbf{S}_n^t)||_2 + ||\mathbf{dir}(\mathbf{S}_n^t)||_2 - ||\mathbf{S}_n^t - \mathbf{S}_{n+2}^t||_2)}{||\mathbf{S}_n^t - \mathbf{S}_{n+2}^t||_2^3}}.$$

We optimize all loss terms together to solve $\{\mathbf{S}_n^{t+1}\}$ given $\{\mathbf{S}_n^t\}$ with:

$$\begin{aligned}\mathcal{L}_{hair} = &\mathcal{L}_{of} + \omega_{hdir}\mathcal{L}_{hdir} + \omega_{hpos}\mathcal{L}_{hpos} \\ &+ \omega_{len}\mathcal{L}_{len} + \omega_{tang}\mathcal{L}_{tang} + \omega_{cur}\mathcal{L}_{cur}.\end{aligned}$$

By utilizing momentum information across the temporal axis, we can provide a better initialization of $\mathbf{S}_n^{t+1}$ given its trajectory and intialize $\mathbf{S}_n^{t+1}$ as

$$\mathbf{S}_n^{t+1} = 3\mathbf{S}_n^t - 3\mathbf{S}_n^{t-1} + \mathbf{S}_n^{t-2}.$$

3.2. HVH

Background. Similar to MVP, we define volumetric primitives $\mathcal{V}_n = \{\mathbf{t}_n, \mathbf{R}_n, \mathbf{s}_n, \mathbf{V}_n\}$ to model a volume of local 3D space each, where $\mathbf{R}_n \in SO(3), \mathbf{t}_n \in \mathbb{R}^3$ describes the volume-to-world transformation, $\mathbf{s}_n \in \mathbb{R}^3$ are the per-axis scale factors and $\mathbf{V}_n = [\mathbf{V}_c, \mathbf{V}_\alpha] \in \mathbb{R}^{4\times M\times M\times M}$ is a volumetric grid that stores three channel color and opacity information. The volumes are placed on a UV-map that are unwrapped from a head tracked mesh and are regressed from a 2D CNN. Using an optimized BVH implementation, we can efficiently determine how the rays intersect each volume and find hit boxes. For each ray $\mathbf{r}_p(t) = \mathbf{o}_p + t\mathbf{d}_p$, we denote (t_{min}, t_{max}) as the start and end point for ray integration. Then, the differentiable aggregation of those volumetric primitives is defined as:

$$\begin{aligned}\mathcal{I}_p &= \int_{t_{min}}^{t_{max}} \mathbf{V}_c(\mathbf{r}_p(t))\frac{dT(t)}{dt}dt, \\ T(t) &= min(\int_{t_{min}}^{t} \mathbf{V}_\alpha(\mathbf{r}_p(t))dt, 1).\end{aligned}$$

We composite the rendered image as $\tilde{\mathcal{I}}_p = \mathcal{I}_p + (1 - \mathcal{A}_p)I_{p,bg}$ where $\mathcal{A}_p = T(t_{max})$ and $I_{p,bg}$ is the background image.

Encoder. The encoder uses the driving signal of a specific point in time and outputs a global latent code $\boldsymbol{z} \in \mathbb{R}^{256}$. We use the tracked guide hairs $\{\mathbf{S}_n^t\}$ and tracked head mesh vertices $\{\boldsymbol{v}_m^t\}$ to define the driving signal. Symmetrically, we learn another decoder in parallel with the encoder in an auto-encoding way that regresses the tracked guide hairs $\{\mathbf{S}_n^t\}$ and head mesh vertices $\{\boldsymbol{v}_m^t\}$ from the global latent code $\boldsymbol{z}$. The architecture of the encoder is an MLP that regresses the parameter of a normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}), \boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathcal{R}^{256}$. We use the reparameterization trick from [17] to sample $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ in a differentiable way.

Hair Volume Decoder. Besides the volumes that are attached to the tracked mesh $\{\boldsymbol{v}_m^t\}$, we define additional hair volume $\mathcal{V}_n^{hair}$ that are associated with guide hair nodes $\mathbf{S}_n^t$. The position $\mathbf{t}_n = \hat{\mathbf{t}}_n + \delta_{\mathbf{t}_n}$, orientation $\mathbf{R}_n = \delta_{\mathbf{R}_n} \cdot \hat{\mathbf{R}}_n$ and scale $\mathbf{s}_n = \hat{\mathbf{s}}_n + \delta_{\mathbf{s}_n}$ of each hair volume are determined by the base hair transformation $(\hat{\mathbf{t}}_n, \hat{\mathbf{R}}_n, \hat{\mathbf{s}}_n)$ and regressed hair relative transformation $(\delta_{\mathbf{t}_n}, \delta_{\mathbf{R}_n}, \delta_{\mathbf{s}_n})$. The base translation $\hat{\mathbf{t}}_n$ of each hair node is directly its position $\mathbf{S}_n^t$. The base rotation $\hat{\mathbf{R}}_n$ is derived from the hair tangential direction and the hair-head relative position. We denote τ_n as the hair tangential direction at position $\mathbf{S}_n^t$ and ν'_n as the direction pointing to the tracked head center starting from S_n^t. Then, the base rotation is $\hat{\mathbf{R}}_n = [\tau_n^T; \rho_n^T; \nu_n^T]$, where $\rho_n = \tau_n \times \nu'_n, \nu_n = \rho_n \times \tau_n$.

The geometry of hair can not be simply described by a surface. Therefore, we design a 2D CNN that convolves along the hair growing direction and the rough hair spatial position separately. Specifically, in the each layer of the 2D CNN, we seperate a $k \times k$ filter into two $k \times 1$ and $1 \times k$ filters and apply convolution along two orthogonal directions respectively, similar to [70]. To learn a more consistent hair shape and appearance model, we optimize per-strand hair features $\{f_n^t\}$ that are shared across all time frames besides the temporally varying global latent code $\boldsymbol{z}$. For each node $\mathbf{S_n^t}$ on a hair strand $\mathbf{S}^t$, we assign an unique feature vector f_n^t. The shared per-strand hair features and the temporal varying latent code $\boldsymbol{z}$ are fused to serve as the input to the hair volume decoder, which is shown in Fig. 2.

Differentiable Volumetric Raymarching of 3D Scene Flow. Learning a volumetric scene representation by multi-view photometric information is sufficient for high fidelity rendering and novel view synthesis. However, it is challenging for the model to reason about motion given the limited supervision and the results have poor temporal consistency, especially on unseen sequences. To better enforce temporal consistency, we develop a differentiable volumetric ray marching algorithm of 3D scene flow which enables training via multi-view 2D optical flow.

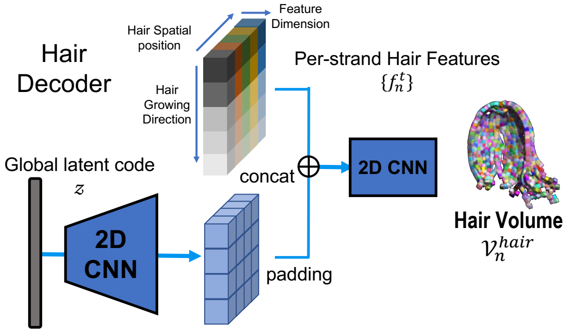

Figure 2. **Architecture of the hair decoder.** The hair decoder takes both the global latent code z and the per-strand hair features $\{f_n^t\}$ as inputs. z is first deconvolved into a 2D feature tensor. It is then padded and concatenated with $\{f_n^t\}$. In the following operation, the 2D convolution layers are applied along the hair growing direction and the hair spatial position seperately.

Given the transformations of each primitive as $(\mathbf{t}_n, \mathbf{R}_n, \mathbf{s}_n)$, we express the coordinate of each node on a volumetric grid at frame u as $\mathbf{V}^u_{xyz} = \mathbf{s}_t \mathbf{R}_t \mathbf{V}_{tpl} + \mathbf{t}_n$, where $\mathbf{V}_{tpl}$ are the coordinates of a 3D mesh grid ranging between $[-1, 1]$. Given that the 3D scene flow from frame u to $u + \delta$ can be expressed by each volumetric primitives as $\{\delta \mathbf{V}^{u,u+\epsilon}_{xyz} = \mathbf{V}^{u+\epsilon}_{xyz} - \mathbf{V}^u_{xyz}\}$ and rendered into 2D flow as:

$$\mathcal{I}^{u,u+\delta}_{p,flow} = \int_{t_{min}}^{t_{max}} (\delta \mathbf{V}^{u,u+\epsilon}_{xyz}(\mathbf{r}_p(t))) \frac{dT(t)}{dt} dt,$$
$$T(t) = min(\int_{t_{min}}^{t_{max}} \mathbf{V}^u_\alpha(\mathbf{r}_p(t))dt, 1).$$

Training Objectives. We train our model in an end-to-end manner with the following loss:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{pho} + \lambda_{flow}\mathcal{L}_{flow} + \lambda_{geo}\mathcal{L}_{geo} \\ &+ \lambda_{vol}\mathcal{L}_{vol} + \lambda_{cub}\mathcal{L}_{cub} + \lambda_{KL}\mathcal{L}_{KL}.\end{aligned}$$

The first term $\mathcal{L}_{pho}$ is the photometric loss that compares the difference between the rendered image $\tilde{\mathcal{I}}_p$ and ground truth image I_p on all sampled pixels $p \in \mathcal{P}$,

$$\mathcal{L}_{pho} = \sum_{p \in \mathcal{P}} ||I_{p,gt} - \tilde{\mathcal{I}}_p||^2_2.$$

The second term $\mathcal{L}_{flow}$ aims to enforce temporal consistency of volumetric primitives from frame u and its adjacent frame $u + \epsilon$ by minimizing the projected 2D flow and ground truth optical flow $I^{u,u+\epsilon}_{p,flow}$,

$$\mathcal{L}_{flow} = \sum_{p \in \mathcal{P}} \mathcal{A}_p ||I^{u,u+\epsilon}_{p,of} - \mathcal{I}^{u,u+\epsilon}_{p,flow}||^2_2,$$

where $\epsilon \in \{-1, 1\}$. It is important to note that we use $\mathcal{A}_p$ to mask out the background part and we do not back propagate the errors from $\mathcal{L}_{flow}$ to $\mathcal{A}_p$ in order to get rid of the background noise in optical flows. To better enforce hair and head primitives moving with the tracked head mesh and guide hair strands, $\mathcal{L}_{geo}$ is designed to measure the difference between the mesh/strand vertices and their corresponding regressed value.

$$\mathcal{L}_{geo} = \sum_n ||\boldsymbol{S}^t_n - \boldsymbol{S}^t_{n,gt}||^2_2 + \sum_m ||\boldsymbol{v}^t_m - \boldsymbol{v}^t_{m,gt}||^2_2,$$

where $\boldsymbol{S}^t_n$ and $\boldsymbol{v}^t_m$ are the coordinate of the nth node of the tracked guide hair and tracked head mesh at frame t and the X_{gt} denotes the corresponding ground truth value.

We also add several regularization terms to inform the layout of the volumetric primitives:

$$\mathcal{L}_{vol} = \sum_{i=1,\cdots,N_p} \prod_{j \in \{x,y,z\}} s^j_i,$$
$$\mathcal{L}_{cub} = \sum_{i=1,\cdots,N_p} ||max(s^x_i, s^y_i, s^z_i) - min(s^x_i, s^y_i, s^z_i)||,$$

where N_p stands for the total number of volumetric primitives and s^x_i, s^y_i, s^z_i are the three entries of each volumetric primitive's scale $\boldsymbol{s}_j$. The two regularization terms aim to prevent each primitive from growing too big while preserving the aspect ratio so that they remain approximately cubic. The last term is the Kullback-Leibler divergence loss $\mathcal{L}_{KL}$ which makes the learnt distribution of latent code $\boldsymbol{z}$ smooth and enforces similarity with a normal distribution $\mathcal{N}(0, 1)$.

4. Experiments

4.1. Dataset

For each video recorded with our multi camera system, we split the them by the motions performed (like nodding and shaking of the head) and hold out the last $\frac{1}{4}$ of each motion for testing drivable animation. This results in roughly 300 frames for training sequence and 100 frames for testing sequence. Additionally, on the training sequence, we hold out 7 cameras that are distributed around the rear and side view of the head. The captured images are downsampled to 1024×667 resolution for training and testing. We train our model exclusively on the training portion of each sequence with $m = 93$ training views.

4.2. Novel View Synthesis

We show both qualitative and quantitative comparisons with other methods [22, 27, 59] on the novel view synthesis task. In the left of Tab. 1, we show the mean squared error (MSE), SSIM and PSNR between predicted images

	Seq01			Seq02			Seq03		
	MSE	SSIM	PSNR	MSE	SSIM	PSNR	MSE	SSIM	PSNR
PFNeRF	51.25	0.9269	31.16	103.41	0.8659	28.15	76.59	0.9000	29.50
NSFF	50.13	0.9346	31.21	90.06	0.8885	28.75	83.18	0.8936	29.1
NRNeRF	56.78	0.9231	30.78	132.16	0.8549	27.13	79.83	0.8987	29.33
MVP	47.54	0.9476	31.6	77.23	0.9088	29.62	73.78	0.9224	29.66
Ours	**41.89**	**0.9543**	**32.17**	**59.84**	**0.9275**	**30.69**	**71.58**	**0.9314**	**29.81**

	Seq01			Seq02			Seq03		
	MSE	SSIM	PSNR	MSE	SSIM	PSNR	MSE	SSIM	PSNR
MVP	47.54	0.9476	31.6	77.23	0.9088	29.62	73.78	0.9224	29.66
MVP w/ $\mathcal{L}_{flow}$	46.49	0.9473	31.69	71.07	0.9107	29.93	75.13	0.9240	29.58
Ours w/o $\mathcal{L}_{flow}$	43.82	0.9508	31.99	65.98	0.9186	30.27	69.97	0.9359	29.93
Ours	**41.89**	**0.9543**	**32.17**	**59.84**	**0.9275**	**30.69**	**71.58**	**0.9314**	**29.81**
MVP	75.68	0.9200	29.49	85.10	0.9039	29.62	83.76	0.9086	29.16
MVP w/ $\mathcal{L}_{flow}$	67.86	0.9276	30.00	83.11	0.9037	29.93	80.96	0.9086	29.16
Ours w/o $\mathcal{L}_{flow}$	71.90	0.9223	29.74	72.74	0.9137	30.27	78.34	0.9198	29.44
Ours	**65.96**	**0.9280**	**30.09**	**67.75**	**0.9208**	**30.69**	**75.66**	**0.9222**	**29.57**

Table 1. **Novel view synthesis**. On the left, we compare our method with both NeRF stemmed methods like NSFF [22], NRNeRF [59] and a per-frame NeRF (PFNeRF) baseline, and a volumetric method like MVP [27]. On the right, we further compare our method and different variants of our methods with MVP on novel views of both seen (top) and unseen (bottom) sequences.

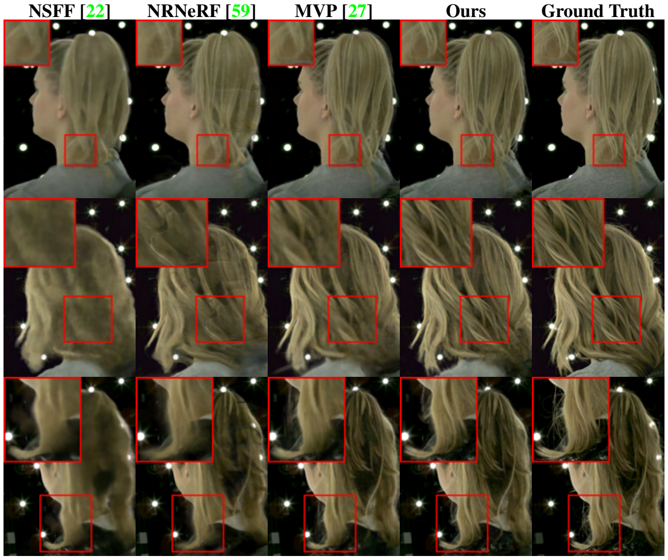

Figure 3. **Comparison on novel view synthesis between different methods.** Please see supplementary material for a bigger version of this figure.

and ground truth images from the novel views of the training sequences. Qualitative results are shown in Fig. 3. Our method has smaller image prediction errors and is able to generate sharper results, especially on the hair regions.

4.3. Ablation Studies

Temporal consistency. To test the effects of the temporal consistency and the tracked guide hair, we also conduct a novel view synthesis task on the test portion of our captured sequence. Note that our model is not trained using any part of the test sequence data. On the right of Tab. 1, we report MSE, SSIM, PSNR on novel views of both seen and unseen sequences. As we can see, having the coarse level guide hair strands tracked and without flow supervision gives us better rendering quality. With flow supervision, the results are improved further. This improvement is because the tracking information helps the volumetric primitives to better localize the hair region with higher consistency. While the improvement for seen motions is relatively small, both our model and MVP are notably improved for unseen sequences with novel hair motion when flow supervision is added. Rendering results on unseen sequences are shown in Fig. 4. In Fig. 5, we visualize the volumetric primitives of the hairs of our model with and without flow supervision. Including flow supervision produces notably better disentanglement between the hair and shoulder.

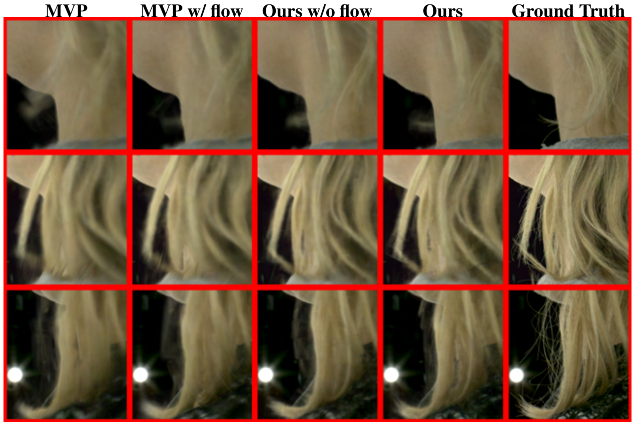

Figure 4. **Ablation of temporal consistency.** We compare our method and MVP w/ and w/o flow supervision. With flow supervision, better temporal consistency and generalization for unseen sequence can be observed. Please see supplementary for a bigger version of this figure.

Hair tracking analysis. We first study the impact of different objectives $\mathcal{L}_{len} + \mathcal{L}_{tang}$ and $\mathcal{L}_{cur}$ in hair tracking. As in Fig. 6, when both $\mathcal{L}_{cur}$ and $\mathcal{L}_{len} + \mathcal{L}_{tang}$ are applied, the tracking results are more smooth and without kinks. We observe that, when using the loss $\mathcal{L}_{len} + \mathcal{L}_{tang}$ as the only regularization term, the length of each hair strand segments are already preserved but could cause some kinks without awareness of the correct hair strand curvatures. $\mathcal{L}_{cur}$ itself does not help and exaggerates the error when the hair strand length is not correct, but yields smooth results when combined with $\mathcal{L}_{len} + \mathcal{L}_{tang}$. This is because curvature computation is agnostic to absolute length of the hair and only controls the relative length ratio.

We show the impact of different initialization for hair tracking in Fig. 7. When no momentum information from

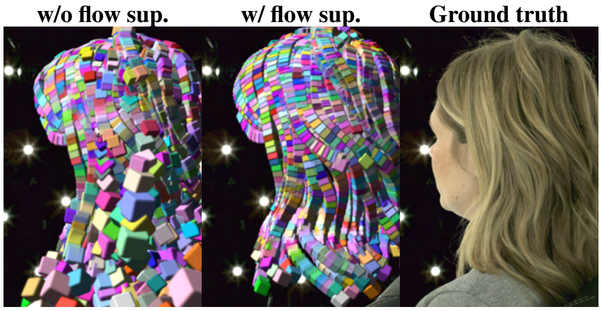

Figure 5. **Ablation on flow supervision.** We further compare the volumetric primitives of the models w/ and w/o flow supervision. We see that model with additional flow supervision yields a consistent and reasonable shape for hair and yields better hair shoulder disentanglement.

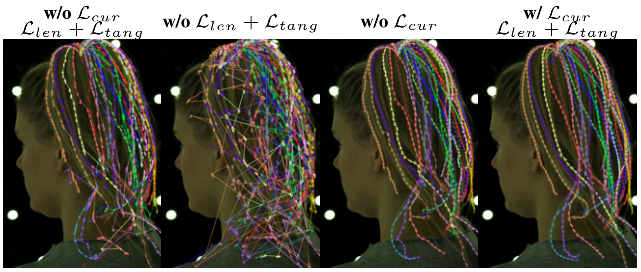
w/o $\mathcal{L}_{cur}$
$\mathcal{L}_{len} + \mathcal{L}_{tang}$
w/o $\mathcal{L}_{len} + \mathcal{L}_{tang}$
w/o $\mathcal{L}_{cur}$
w/ $\mathcal{L}_{cur}$
$\mathcal{L}_{len} + \mathcal{L}_{tang}$

Figure 6. **Effects of $\mathcal{L}_{len} + \mathcal{L}_{tang}$ and $\mathcal{L}_{cur}$.** We show how the shape and curvature of tracked hair strands are preserved with both $\mathcal{L}_{len} + \mathcal{L}_{tang}$ and $\mathcal{L}_{cur}$.

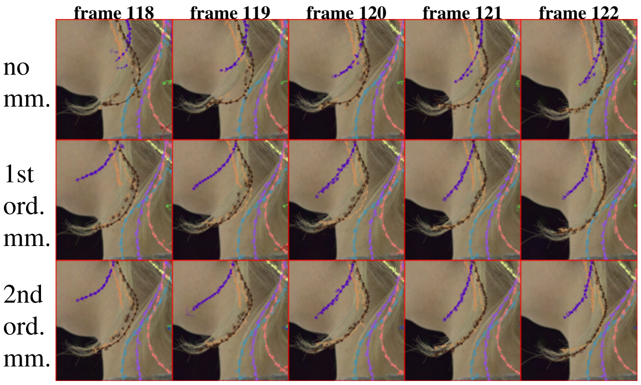

Figure 7. **Ablation of different initialization in hair tracking.** We show tracking results of our methods with different initializations. From top to bottom, we use no momentum information, first and second order momentum information for tracking initialization. Please note the brown and orange strands. As we can see, the hairs are better tracked when we utilize the dynamic information from previous frames. Better view in color version.

previous frames is used, there is more obvious drifting on some of the strands happening, while the drifting is less severe when we take advantage of the motion information from previous frames.

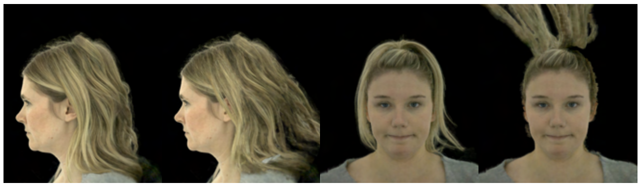

Figure 8. **Hair position editing.** We create a new animation by direct editing on the guide hair strands. As we can see the volumes of hair are driven by the lifted guide hair to create a new hair motion. Please see supplementary material for video results.

5. Applications and Limitations

One major application that is enabled by our neural volumetric scene representation is novel view synthesis as we have shown in Sec. 4.2. Our neural volumetric representation is also animatable with a sparse driving signals like guide hair strands. Given that we have explicitly modeled hair in the form of guide strands, our method allows modifying the guide hairs directly. In Fig. 8, we show four snapshots of different configurations of hair positions. Please see more results and details in the supplementary material.

There are several limitations of our work which we plan to address in the future: 1) Our method requires the help from artist to prepare guide hair at the first frame and some flyaway hair might be excluded. 2) We currently do not consider physics based interactions between hair and other objects like the shoulder or the chair. 3) Although we achieved certain level of disentanglement between hair and other objects without any human labeling, it is still not perfect. We only showed results on blonde hair which could be better distinguished from a dark background. Our method might be limited by other hairstyles like buzz cut or afro-textured that are hard for artist to prepare guide hair. Please see the supplementary material for more discussions on hair styles and generalization. Future directions like incorporating a physics aware module or leveraging additional supervision from semantic information for disentanglement could be interesting.

6. Discussion

In this paper, we present a hybrid neural volumetric representation for hair dynamic performance capture. Our representation leverages the efficiency of guide hair representation in hair simulation by attaching volumetric primitives to them as well as the high DoF of volumetric representation. With both hair tracking and 3D scene flow refinement, our model enjoys better temporal consistency. We empirically show that our method generates sharper and higher quality results on hair and our method achieves better generalization. Our model also supports multiple applications like drivable animation and hair editing.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European conference on computer vision*, pages 696–712. Springer, 2020. 1, 2

[2] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *European Conference on Computer Vision*, pages 441–459. Springer, 2020. 1, 2

[3] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 1, 2

[4] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 4

[5] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020. 1, 2

[6] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, pages 608–625. Springer, 2020. 3

[7] Menglei Chai, Changxi Zheng, and Kun Zhou. Adaptive skinning for interactive hair-solid simulation. *IEEE transactions on visualization and computer graphics*, 23(7):1725–1738, 2016. 3

[8] Ricson Cheng, Ziyan Wang, and Katerina Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2

[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016. 2

[10] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 3

[11] Liwen Hu, Derek Bradley, Hao Li, and Thabo Beeler. Simulation-ready hair capture. In *Computer Graphics Forum*, volume 36, pages 281–294. Wiley Online Library, 2017. 2

[12] Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. Robust hair capture using simulated examples. *ACM Transactions on Graphics (TOG)*, 33(4):1–10, 2014. 2

[13] Hayley Iben, Mark Meyer, Lena Petrovic, Olivier Soares, John Anderson, and Andrew Witkin. Artistic simulation of curly hair. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 63–71, 2013. 3

[14] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 3

[15] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[16] Tero Karras and Timo Aila. Fast parallel construction of high-quality bounding volume hierarchies. In *Proceedings of the 5th High-Performance Graphics Conference*, pages 89–99, 2013. 3

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[18] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast optical flow using dense inverse search. In *European Conference on Computer Vision*, pages 471–488. Springer, 2016. 4

[19] Christoph Lassner and Michael Zollhöfer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021. 1, 2

[20] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194–1, 2017. 1, 2

[21] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *arXiv preprint arXiv:2103.02597*, 2021. 1, 3

[22] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 1, 3, 6, 7

[23] David B Lindell, Julien NP Martel, and Gordon Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14556–14565, 2021. 3

[24] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15651–15663. Curran Associates, Inc., 2020. 1, 3

[25] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4), July 2018. 1, 2

[26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images.

ACM Transactions on Graphics (TOG), 38(4), July 2019. 1, 2, 3

[27] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4), July 2021. 1, 3, 6, 7

[28] 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. 3

[29] Linjie Luo, Hao Li, Sylvain Paris, Thibaut Weise, Mark Pauly, and Szymon Rusinkiewicz. Multi-view hair capture using orientation fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1490–1497. IEEE, 2012. 2

[30] Linjie Luo, Hao Li, and Szymon Rusinkiewicz. Structure-aware hair capture. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013. 2

[31] Linjie Luo, Cha Zhang, Zhengyou Zhang, and Szymon Rusinkiewicz. Wide-baseline hair capture using strand-based refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 265–272, 2013. 2

[32] 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. 3

[33] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6878–6887, 2019. 1, 2

[34] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1, 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. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 3

[36] Giljoo Nam, Chenglei Wu, Min H Kim, and Yaser Sheikh. Strand-accurate multi-view hair capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 155–164, 2019. 2, 4

[37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 3

[38] Sylvain Paris, Hector M Briceno, and François X Sillion. Capture of hair geometry from multiple images. *ACM Transactions on Graphics (TOG)*, 23(3):712–719, 2004. 2

[39] Sylvain Paris, Will Chang, Oleg I Kozhushnyan, Wojciech Jarosz, Wojciech Matusik, Matthias Zwicker, and Frédo Durand. Hair photobooth: geometric and photometric acquisition of real hairstyles. *ACM Transactions on Graphics (TOG)*, 27(3):30, 2008. 2

[40] 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. 3

[41] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1, 3

[42] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 1, 3

[43] 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. 3

[44] 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, 3

[45] Lena Petrovic, Mark Henne, and John Anderson. Volumetric methods for simulation and rendering of hair. *Tech. Rep. Pixar Animation Studios*, 2(4), 2005. 3

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1, 3

[47] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 1, 3

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[49] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *arXiv preprint arXiv:2110.06635*, 2021. 1, 2

[50] 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. 3

[51] 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. 3

[52] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 1

[53] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2, 3

[54] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[55] Tiancheng Sun, Giljoo Nam, Carlos Aliaga, Christophe Hery, and Ravi Ramamoorthi. Human Hair Inverse Rendering using Multi-View Photometric data. In Adrien Bousseau and Morgan McGuire, editors, *Eurographics Symposium on Rendering - DL-only Track*. The Eurographics Association, 2021. 2, 4

[56] Richard Szeliski and Polina Golland. Stereo matching with transparency and matting. In *International Conference on Computer Vision*, pages 517–524. IEEE, 1998. 1, 2

[57] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 1, 2

[58] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018. 1, 2

[59] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 1, 3, 6, 7

[60] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2634, 2017. 2

[61] Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2595–2603, 2019. 2

[62] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 3

[63] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, June 2021. 1, 3

[64] Yichen Wei, Eyal Ofek, Long Quan, and Heung-Yeung Shum. Modeling hair from multiple views. In *ACM SIGGRAPH 2005 Papers*, pages 816–820. 2005. 2

[65] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 1, 2

[66] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 82–90, 2016. 2

[67] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 1, 3

[68] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 1, 2

[69] Zexiang Xu, Hsiang-Tao Wu, Lvdi Wang, Changxi Zheng, Xin Tong, and Yue Qi. Dynamic hair capture using spacetime optimization. *ACM Transactions on Graphics (TOG)*, 33(6), nov 2014. 2

[70] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5

[71] Lingchen Yang, Zefeng Shi, Youyi Zheng, and Kun Zhou. Dynamic hair modeling from monocular videos using deep neural networks. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019. 2

[72] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2492–2502. Curran Associates, Inc., 2020. 3

[73] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 3

[74] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruc-

tion of rigid objects in motion with neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13144–13152, 2021. 1, 3

[75] Qing Zhang, Jing Tong, Huamin Wang, Zhigeng Pan, and Ruigang Yang. Simulation guided hair dynamics modeling from video. In *Computer Graphics Forum*, volume 31, pages 2003–2010. Wiley Online Library, 2012. 2

[76] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4), July 2018. 1, 2

[77] Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, and Simon Lucey. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 57–65, 2017. 2