HairOrbit: Multi-view Aware 3D Hair Modeling from Single Portraits

Leyang Jin[1], Yujian Zheng[1⋆], Bingkui Tong[1], Yuda Qiu[2], Zhenyu Xie[1], and Hao Li[1,3]

[1] Mohamed bin Zayed University of Artificial Intelligence
[2] SSE, The Chinese University of Hong Kong, Shenzhen
[3] Pinscreen

Abstract. Reconstructing strand-level 3D hair from a single-view image is highly challenging, especially when preserving consistent and realistic attributes in unseen regions. Existing methods rely on limited frontal-view cues and small-scale/style-restricted synthetic data, often failing to produce satisfactory results in invisible regions. In this work, we propose a novel framework that leverages the strong 3D priors of video generation models to transform single-view hair reconstruction into a calibrated multi-view reconstruction task. To balance reconstruction quality and efficiency for the reformulated multi-view task, we further introduce a neural orientation extractor trained on sparse real-image annotations for better full-view orientation estimation. In addition, we design a two-stage strand-growing algorithm based on a hybrid implicit field to synthesize the 3D strand curves with fine-grained details at a relatively fast speed. Extensive experiments demonstrate that our method achieves state-of-the-art performance on single-view 3D hair strand reconstruction on a diverse range of hair portraits in both visible and invisible regions.

Keywords: 3D Hair Modeling · Single-view Reconstruction · Neural Network

1 Introduction

Reconstructing strand-level 3D hair geometry from a single image remains one of the most challenging tasks in digital human modeling. Hair exhibits highly complex structures with self-occlusions, intricate topology, and diverse local details such as curliness, partition, and varying lengths. In single-view reconstruction, which is an inherently ill-posed problem, the visible regions provide only partial cues, while the occluded parts must be inferred from learned priors that ensure both geometric plausibility and stylistic consistency. Achieving such coherence between visible and invisible regions is essential for realistic digital avatars and AR/VR applications, yet it remains an open challenge.

Existing single-view methods often rely on limited appearance cues and small-scale or style-restricted 3D hair datasets [25, 30, 35, 42] to hallucinate occluded

⋆ Corresponding author.

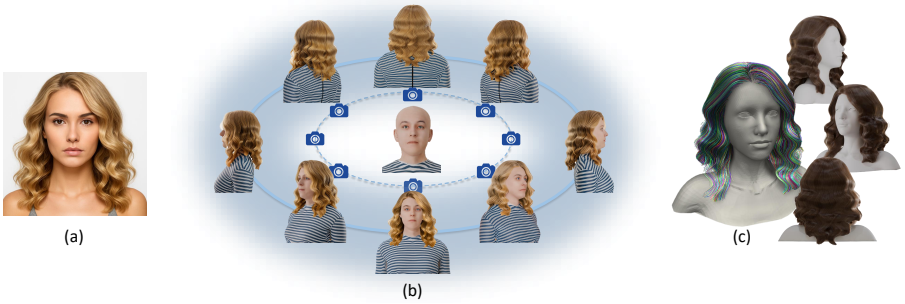

Fig. 1: We propose a novel framework for strand-level single-view 3D hair reconstruction. Given a frontal-view portrait (a), we first synthesize corresponding calibrated multi-view images (b) on a camera orbit, then reconstruct multi-view aware 3D hair strands (c). Note that the left view in (c) is rendered with about 10k strands to better visualize the geometry, while the other 3 views are rendered with 100k.

geometry. However, due to the scarcity and limited realism of current synthetic 3D hair datasets [25, 44], these priors are insufficient to capture the diversity and structural coherence of real-world hairstyles, leading to unnatural, over-smoothed, or image-unaligned reconstructions. Notably, a non-negligible limitation of these methods is that the reconstructed hair often appears plausible from the frontal view but fails to maintain style consistency or suffers from severe artifacts under unseen viewpoints. In contrast, recent state-of-the-art multi-view approaches [9, 28, 34, 37, 45] demonstrate impressive performance in capturing high-fidelity, strand-level details, as they can leverage fine and geometrically consistent information from captured multi-view images. This raises a natural question: can multi-view information be synthesized to assist single-view reconstruction?

Inspired by recent advancements in video diffusion models [33], we observe that such models intrinsically encode strong 3D priors through their temporal consistency. Building on this insight, we combine the priors learned from large-scale video generation models with a limited but expertly curated set of high-quality synthetic 3D hair data [7,25], to validate our hypothesis that single-view hair reconstruction can be reformulated as a multi-view task. Specifically, we fine-tune a video diffusion model using a lightweight LoRA [5] on a carefully selected subset of 3D hair models and their corresponding multi-view renderings, enabling the generation of consistent video sequences in a 360° orbital rotation from a single input image. By leveraging these synthesized multi-view sequences as an intermediate bridge, we effectively transform the single-view reconstruction problem into a calibrated multi-view setting, allowing us to exploit the advantages of multi-view methods without requiring captured real multi-view data (Fig. 1).

However, when revisiting multi-view hair strand reconstruction, several issues remain. The most critical one lies in the dependency of current methods

on orientation maps extracted using Gabor filters [21]. Such maps are typically noisy and inaccurate, which severely affects the geometric consistency across views. Existing approaches either rely on time-consuming denoising [34] or optimization processes [37] to achieve satisfactory results, or suffer from degraded quality [9]. To address this, we train a neural orientation extractor using manually annotated hair-direction data from [42], which produces accurate, clean, and smooth orientation maps. Unlike [42], which aims to predict a directed global orientation that requires semantic understanding and therefore struggles to generalize to back or side views when only trained on frontal views, we observe that orientation extraction is essentially a local, pixel-aligned task. Consequently, by training on annotated frontal-view orientations, our model generalizes well to full-view cases. The effectiveness of our method is validated on a newly annotated dataset of 395 images with hair growth strokes similar to [42], but covering full-view cases captured from multi-view videos [28, 34].

Following [9, 34, 35, 42], we predict an implicit 3D hair field to guide strand growth based on the aggregated features from multi-view orientation maps and estimated depth maps. This leads to the second question concerning the implicit field representation and the hair-growing mechanism. Existing methods typically construct two separate fields: an orientation field and an occupancy field. We find this occupancy field actually inefficient for hair growing. Instead, we introduce a hybrid field that jointly models both behaviors: within the hair volume, it learns the orientation field, and outside, it predicts zero-valued growth vectors. With this formulation, hair strands naturally stop when reaching the boundary, and repeated growth steps always converge to the same endpoints. This property enables highly efficient parallel strand growth, leading to a substantial speedup. We combine scalp-root and segment-based growth to achieve fine and efficient strand generation [3], avoiding the incompleteness of HairStep and the high cost of MonoHair. In addition, we filter out redundant or collapsed short strands brought by pre-sampled roots and recover collapsed buzz-cut using a re-projection method.

The main contributions of our work are as follows:

- We reformulate single-view hair reconstruction as a calibrated multi-view reconstruction problem by leveraging the intrinsic cross-view priors embedded in video diffusion models.
- We propose a multi-view neural orientation extractor, a hybrid implicit field, and a combined strand-growing algorithm to effectively balance reconstruction fidelity and efficiency for more practical applications.
- We achieve state-of-the-art performance in single-view strand-level 3D hair reconstruction, simultaneously maintaining visual alignment with the input image and ensuring plausible multi-view geometry.

2 Related Work

Strand-level 3D hair modeling How to acquire human hair in 3D has been an active research topic for decades. Multi-view stereo methods have long been ex-

plored to reconstruct 3D hair geometry from multiple captured images [6,15,16, 18,22,45]. To reduce the reliance on specialized equipment and constrained capture setups, several approaches attempt to model hair from less restricted inputs, such as sparse-view images [38], selfie videos [12], and RGB-D streams [39]. More recently, high-quality reconstructions from monocular videos have been demonstrated [28,34]. However, these methods require long processing times (e.g., 3–4 days in [28] and 4–6 hours in [34]), limiting their accessibility to ordinary users.

In contrast, several studies focus on reconstructing 3D hair from a single-view image, which are more efficient. Early methods [3,4] lift partial strands from 2D to manipulate portraits. The pioneering retrieval-based methods [2,7] typically retrieve a coarse hair model from a database and then refine it through geometric optimization. The effectiveness of these approaches relies on the quality of priors, and the performance is less satisfactory for challenging input. With the rise of deep learning, data-driven methods [26,27,35,42,44] have emerged, leveraging deep neural networks for reconstruction. However, these methods often produce coarse or over-smoothed geometry due to the limited prior from the scale, diversity, and quality of 3D synthetic data. Recent approaches [25,29] further generate hair texture maps in a latent space, though such representations are usually misaligned with the input view. Differentiable-rendering-based methods [30,32] alleviate some limitations on the alignment but still remain slow and struggle to recover unseen back-view regions.

Orientation maps for hair modeling The intricate and tangled structure of human hair makes directional cues an intuitive foundation for 3D reconstruction. Instead of modeling hair geometry directly, many approaches first infer intermediate orientation representations that describe strand flow in either 2D or 3D space. Classical image-based methods estimate 2D orientation maps by convolving the input portrait with a bank of Gabor filters and taking the direction that yields the strongest response [21,22]. Such orientation maps can then be lifted into 3D through multi-view calibration [6,15,16] or supplied to neural networks as auxiliary guidance for predicting volumetric hair structures [35,36,40,44]. However, 2D orientations obtained from local filtering are highly sensitive to noise, which can be mitigated via an additional postprocessing process [15,16,34]. [42] introduces a neural orientation extractor that predicts directional strand maps directly from raw images to resolve growing-direction ambiguity. However, it is limited to frontal-view inputs.

3D hair dataset Lack of large-scale and high-quality 3D data remains a major bottleneck in hair modeling research. Currently, USC-HairSalon [7] is the most widely used open-source dataset, containing 343 hair models across 12 unbalanced categories. [44] augments the data from [7] by randomly combining parts from hair strand models with similar styles, but the resulting data are still limited in both realism and diversity. Besides, 3DHW [2] collects 50K portrait images and reconstructs hair geometry through traditional single-view optimization. Due to the inherent limitations of this algorithm, the resulting 3D models often lack fidelity and structural accuracy. [43] proposes a large-scale dataset of

calibrated multi-view images with 2,396 hairstyles, but without corresponding 3D strands, which limits its applicability to strand-level reconstruction tasks. Recently, [25] builds a dataset of 40K 3D hairstyles with rendered images, yet the variation remains limited, with many repetitive styles and unnatural poses. Since creating high-quality 3D hair models at scale is costly, we seek to move away from heavy reliance on such data and leverage multi-view priors to improve the generalization of single-view reconstruction.

3 Method

3.1 Overview

Aiming to improve the style consistency and overall quality of single-view hair reconstruction, we propose a novel pipeline, *HairOrbit*, as illustrated in Fig. 2. Given a frontal-view portrait I_{in}, we first align the hair region with a template body image $I_{template}$ to obtain $I_{aligned}$ by fitting landmarks as [43]. From $I_{aligned}$, we generate calibrated multi-view images $\{I_m^i\}$ using a video diffusion model (see Sec. 3.2). Based on $\{I_m^i\}$, we reconstruct a 3D mesh M and render the corresponding multi-view depth maps $\{D^i\}$. In parallel, multi-view orientation maps $\{O^i\}$ are predicted from $\{I_m^i\}$. Subsequently, we predict a hybrid implicit field $\boldsymbol{\mathcal{F}}$ from the aggregated features of $\{O^i\}$ and $\{D^i\}$, to synthesize 3D hair strands $\{S^i\}$ (see Sec. 3.3).

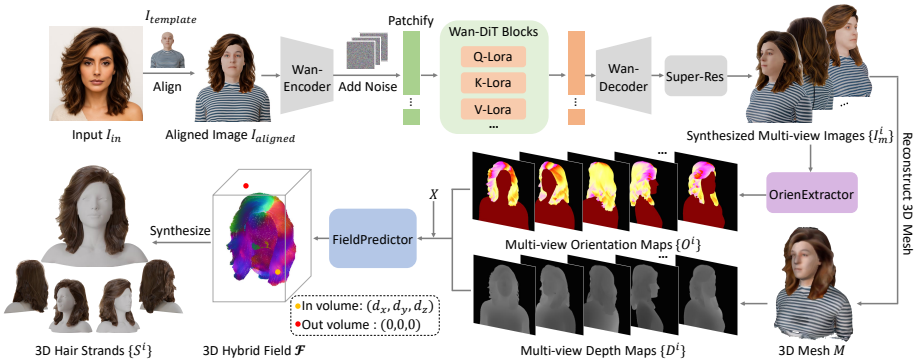

Fig. 2: Overview of *HairOrbit*. Given a single portrait, *HairOrbit* converts single-view 3D hair reconstruction into a multi-view task.

3.2 Multi-view Generation

Given the aligned image $I_{aligned}$, the goal of multi-view generation is to synthesize calibrated multi-view images $\{I_m^i\}$ along a predefined orbital path. We leverage the intrinsic cross-view priors embedded in video diffusion models, and

adapt the model to our setting by training a LoRA [5] on a small set of synthetic multi-view renderings of 3D hairstyles.

Video diffusion model. To ensure temporal consistency in the generated results, we adopt WAN [33] as our backbone. A core element of this framework is a VAE that encodes video frames into a temporally coherent latent representation, ensuring causal consistency while maintaining computational efficiency. In this latent space, a Diffusion Transformer (DiT) is employed to generate latent representations of multi-view 360° rotated video frames from noise, while conditioning on the latent derived from our aligned input image. The generated latents are then decoded into multi-view frames. To accommodate consumer-grade GPU hardware constraints, we employ a lower-resolution version of the model during inference, which leads to synthesized multi-view images that are not sufficiently sharp for accurate orientation extraction. To address this, we enhance the generated multi-view images using the Flux.1-dev Upscaler [4] [10, 11] , obtaining high-resolution results $\{I_m^i\}$ with improved hair texture and fine-grained details.

LoRA training for orbital hair rotation. To guide the video diffusion model toward our desired generation pattern using only a small amount of data, while minimizing the forgetting of its pre-learned real-world hair priors, we fine-tune the model with LoRA applied to the Q, K, V, O projections and the first and last linear layers (FFN.0 and FFN.2) of each transformer block. We prepare multi-view renderings from carefully selected typical 3D hairstyles, covering a wide range of variations in length, curliness, and partition. Following diffusion-based [23] video generation models, we optimize the LoRA parameters using a standard noise prediction loss. Given a clean latent $\mathbf{x}_0$ and a timestep $t \sim \mathcal{U}(1, T)$, a noisy latent $\mathbf{x}_t$ is obtained by the forward diffusion process:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

where $\bar{\alpha}_t$ denotes the cumulative noise schedule. The model $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$ predicts the noise $\boldsymbol{\epsilon}$ conditioned on $\mathbf{c}$, which in our case corresponds to the latent representation of the aligned input image. The training loss is defined as:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\big[\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}\|_2^2\big], \tag{2}$$

where $\mathbf{x}_0$ is the clean latent encoded by the VAE, $\mathbf{x}_t$ is the noisy latent at step t, $\bar{\alpha}_t$ is the cumulative product of noise coefficients, $\boldsymbol{\epsilon}$ is the Gaussian noise to be predicted, ϵ_θ is the DiT backbone with LoRA adapters, and $\mathbf{c}$ is the conditioning latent from the aligned image $I_{aligned}$. During fine-tuning, only the LoRA weights are updated, while all pretrained parameters remain frozen.

3.3 Hair Strands Reconstruction

After the previous stage, we successfully transfer the single-view reconstruction to a calibrated multi-view reconstruction problem. Following [9, 34], we adopt

[4] `https://huggingface.co/jasperai/Flux.1-dev-Controlnet-Upscaler`

the paradigm that predicting an implicit field from multi-view orientation maps and depth maps. Then we generate 3D hair strands according to the field value by certain strand growing algorithm.

Orientation Extraction Due to the inherent difficulty of realistic hair rendering, existing hair modeling approaches [3, 9, 34, 35, 40, 44] typically use orientation maps instead of raw hair images to reduce the domain gap between synthetic and real data. With pixels indicating 2D growing direction in local regions, we hypothesize that providing clean and smooth orientation maps $\{O^i\}$ can significantly benefit 3D reconstruction. However, orientation maps obtained through image filtering are often contaminated by noise and thus fail to serve as reliable intermediate representations, leading to inaccuracies in 3D reconstruction. [34] attempts to address this problem via patch-based multi-view optimization for denoising, but this makes the overall process extremely time-consuming. [42] eliminates the reliance on Gabor filters by training a neural 2D orientation estimator on the HiSa dataset, which provides 2D directional curve annotations, allowing direct prediction of strand orientations from raw images. However, HiSa contains only frontal-view annotations, and the directional strand maps used in [42] rely heavily on global context information. Thus, it often fails on side and back views.

We observe that orientation extraction is inherently a pixel-aligned and locally perceptual task. We utilize frontal-view annotations in HiSa to learn orientation estimation across multiple viewpoints. Following their design, we colorize SVG strand curves to generate training orientation maps within the hair mask region and train a U-Net [24] model using a combination of ℓ_1 and perceptual losses. Although we use the same network as [42], changing the learning target from a directional strand map to an undirectional orientation map, makes our method more applicable and accurate for full-view tasks, rather than being limited to frontal views like HairStep [42].

To verify the performance of the orientation extractor, we also annotate a set of multi-view images with SVG curves and conduct an evaluation on orientation prediction. Please check Sec. 4.2 for details.

Depth estimation Since we have consistent multi-view images $\{I^i_m\}$ from the video diffusion model, it is straightforward for us to reconstruct a mesh M and render depth maps D^i for each frame i. In practice, considering the time efficiency, we use these multi-view images to fit Gaussians [8]. [8] collapses the 3D volume into a set of 2D oriented planar Gaussian disks, providing view-consistent geometry while modeling surfaces intrinsically. After training, a surface mesh can be obtained via marching cubes [14], which will be rendered to multi-view depth.

Hybrid field reconstruction Existing methods typically construct two separate fields: an orientation field and an occupancy field. The occupancy field defines hair boundaries to constrain growth, while the orientation field learns directional cues only within the hair volume and depends on the occupancy field to terminate strands outside. We observe that the occupancy field is inefficient, introducing

costly per-point queries and/or extra ray-casting operations after marching cubes in the hair-growing process. As in DeepSketch2Hair [27], strands grow cell by cell along local directions and stop when exiting the volume, leading to different termination times and preventing full parallelization. Instead, we introduce a hybrid field $\boldsymbol{\mathcal{F}}$ that unifies both behaviors by learning orientations inside the hair volume and predicting zero growth vectors outside.

The implicit function of $\boldsymbol{\mathcal{F}}$ is formulated as:

$$F(X, (O^1, D^1), ..., (O^n, D^n)) = (d_x, d_y, d_z), \tag{3}$$

where the network F takes the predicted 2D orientation maps $\{O^i\}$, rendered depth maps $\{D^i\}$ and a 3D query point X as input to predict its 3D hybrid vector d.

Following [42], we stack $\{O^i\}$ and $\{D^i\}$ of input view i channel-wisely and feed them to a backbone network H [19] to produce a deep feature map $f_i = H(O^i, D^i)$. We then fetch pixel-aligned features $f_i(x)$ of 3D query point X from each view. Then, we concatenate the features from different sources together with the positional encoding [17] of X, and feed them into the decoder to predict the hybrid field value at position X.

The whole network is trained with the average L_1 loss of vector components on each axis. Specifically, we have:

$$L_{\text{ori}} = \frac{1}{N} \sum_{i}^{N} \frac{\|d^* - d\|_1}{3}. \tag{4}$$

Intrinsically, the hybrid field is defined as the product of occupancy and orientation, and this simple reformulation makes strand growth fully parallelizable for a fixed number of steps without boundary checks. With this design, strands naturally stop at the boundary, and repeated growth steps consistently converge to the same endpoints, enabling efficient parallel strand growth and yielding a significant speedup. After the growth process, we can perform a single resampling step to obtain uniformly distributed points along each strand.

Strand growing Similar to the prior method [42], directly growing hair strands $\{S^i\}$ from the scalp roots in our hybrid field is fast but not always complete. When the predicted orientations are slightly inaccurate, the growth direction may deviate near the boundary, causing strands to stop prematurely and resulting in missing tails or unreachable regions. In contrast, the segment-based strategy of [34] grows numerous short strand segments from random seeds within the occupancy field and then connects them using robust geometric rules, yielding highly accurate yet computationally expensive results for full hairstyles.

We observe that scalp-rooted growth already covers the majority of visible regions, leaving only small gaps. Therefore, we adopt a hybrid growing strategy [3]: we first grow the main strands from the scalp, and then generate supplementary short segments only in the missing areas to fill incomplete regions. We attach each supplementary segment to the nearest scalp strand by appending

all the preceding points of the closest strand and smoothing the connection for continuity. This lightweight correction adds negligible computational cost while producing more complete and visually coherent hair reconstructions.

Some pre-sampled roots on the canonical head produce redundant or collapsed short strands. To preserve valid buzz-cut hairs while keeping the overall hair clean, we keep only short strands projected inside the hair masks of at least two views. As the reconstructed hybrid fields may be inaccurate in the buzz-cut region due to its extremely thin geometric structure, the generated buzz-cut hair strands tend to be sparse, and some strands collapse near their roots. To address this issue, we further recover the collapsed roots by reprojecting them into visible views, estimating the mean 3D orientation tangent to the mesh surface across multiple views, and regrowing the strands along the recovered direction.

4 Experiment

4.1 Datasets

To train the LoRA module, we construct a high-quality synthetic multi-view hair dataset. Specifically, representative 3D hair strand models are carefully selected from USC-HairSalon [7] and Difflocks [25]. Each model is registered to a template body mesh and rendered from 97 viewpoints along an orbital path under a static studio lighting setup to form 24 fps videos, with color augmentations for diversity. An additional 100 3D models are randomly selected and rendered for evaluation.

For training the 3D hybrid field, we use 543 hair models in total, including 200 short and curly models from Difflocks and 343 models from USC-HairSalon, together with their mirrored versions with respect to the Y–Z plane. We reserve 5% of the data for evaluation.

To train the orientation extractor, we use 1,250 annotated frontal-view images and apply geometric transformations (translation, scaling, flipping, and rotation) to augment the data following [42]. Additionally, we manually annotate 395 frames from real-world multi-view hair capture videos [28, 34], covering a wide range of hairstyles, to evaluate full-view performance.

4.2 Evaluation

Multi-view generation We evaluate the effectiveness of our multi-view generation pipeline on a selected synthetic dataset, comprising 50 samples from USC-HairSalon [7] and 50 from Difflocks [25]. We first measure the difference between the synthesized novel-view hair image and the rendering of the ground-truth strands using four metrics: L_1 error, PSNR, LPIPS [41]. We also compute the CLIP similarity [13, 31] between the generated views and the input view. The quantitative results are **0.049** for L_1, **20.897** for PSNR, and **0.162** for LPIPS. These values demonstrate that our generative module synthesizes perceptually reasonable novel-view hair. Furthermore, a CLIP similarity score of **0.834** indicates high perceptual consistency between the synthesized view and the input single view. Please check visual results in the supplementary material.

Table 1: Quantitative comparisons of full-view orientation extraction (mean angle error, lower is better). #Views means the number of annotated views.

	nastya	ksyusha	jenya	shortCurly	whiteCurly	midWavy	midCurly	Overall
#Views	62	66	70	58	42	33	64	**395**
Gabor	11.88	16.56	18.88	18.87	24.56	11.31	10.79	**16.05**
HairStep	6.28	10.98	13.16	13.68	21.57	6.32	5.54	**10.88**
Ours	**1.79**	**4.97**	**6.09**	**6.72**	**13.50**	**2.15**	**2.12**	**5.13**

3D hybrid field reconstruction Although we only predict the hybrid field, the occupancy can be easily inferred by checking whether the magnitude of the orientation vector at a query point is sufficiently small (we set a threshold of 0.1 to determine the outside region), since the orientations inside the volume are learned with unit length. To evaluate the predicted fields, we follow the metrics used in [42]. For orientation accuracy, we compute the mean squared error between the predicted and ground-truth orientations over 10k sampled points, achieving a value of **0.018**. For occupancy accuracy, we calculate the IoU and precision, obtaining **0.986** and **0.997**, respectively.

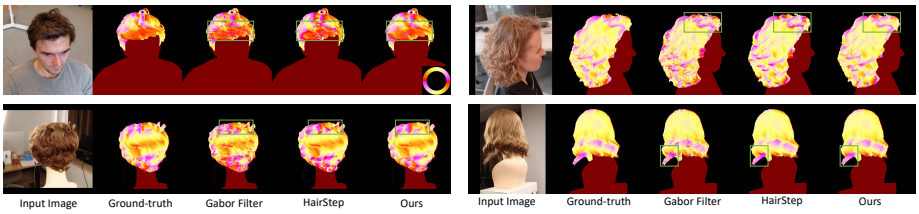

Fig. 3: Qualitative comparisons of full-view orientation extraction. Results of HairStep have been converted to orientation maps.

Full-view orientation extraction Furthermore, we evaluate the performance of our orientation extraction module on the annotated full-view dataset containing 395 images, and compare it against the Gabor filter [44] and the strand map from HairStep [42] (converted into an orientation map for fair comparison). We compute the mean angular error between the predicted and ground-truth orientation maps over all pixels within the hair mask region. As shown in Tab. 1, our extracted orientations significantly outperform both the Gabor filter and the orientation derived from HairStep. A visual comparison is provided in Fig. 3. While we employ the same network architecture as [42], our key difference lies in redefining the learning target from a directional strand map to an undirectional orientation map. This reformulation enables more accurate and generalizable full-view orientation estimation, overcoming the frontal-view limitation of

HairStep [42]. It opens the possibility of removing the reliance of hair modeling on the noisy outputs generated by Gabor filters.

4.3 Comparison

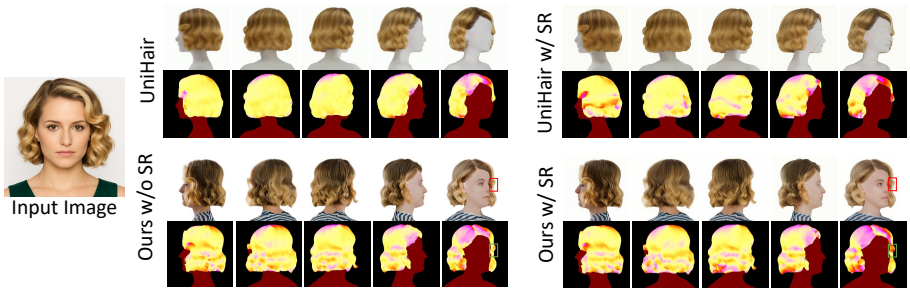

Fig. 4: Comparisons on multi-view generation.

To comprehensively assess the performance of *HairOrbit*, we conduct comparisons on multi-view generation and single-view 3D strands reconstruction, respectively.

Comparison on multi-view generation A recent work, UniHair [43], also attempts to generate multi-view hair images using a diffusion prior fine-tuned on synthetic data with both view-level and pixel-level Gaussian refinements. We compare our multi-view generation module against UniHair and its enhanced version, UniHair-SR, which incorporates a post-processing step using the same super-resolution we adopt. As illustrated in Fig. 4, all methods are capable of generating perceptually reasonable surrounding views from a single input hair image. However, our results exhibit richer real-world hair texture details and are more consistent in style with the input image. In contrast, the results of UniHair appear more synthetic. Furthermore, the hair images generated by UniHair suffer from severe blurring, resulting in inaccurate orientation prediction. Even after super-resolution enhancement, UniHair-SR produces unnatural strand details, which are detrimental to subsequent orientation extraction. In comparison, our multi-view generation module produces high-quality surrounding-view images, significantly improving the performance of orientation extraction. The extracted orientation maps not only capture global hairstyle characteristics, such as overall curl patterns, but also maintain smooth and continuous local details. More results are provided in the supplementary.

Comparison on single-view 3D strands reconstruction We compare our method with recent works HairStep [42], Im2Haircut [30], and Difflocks [25], the state-of-the-art methods on single-view 3D hair strands reconstruction. To compare

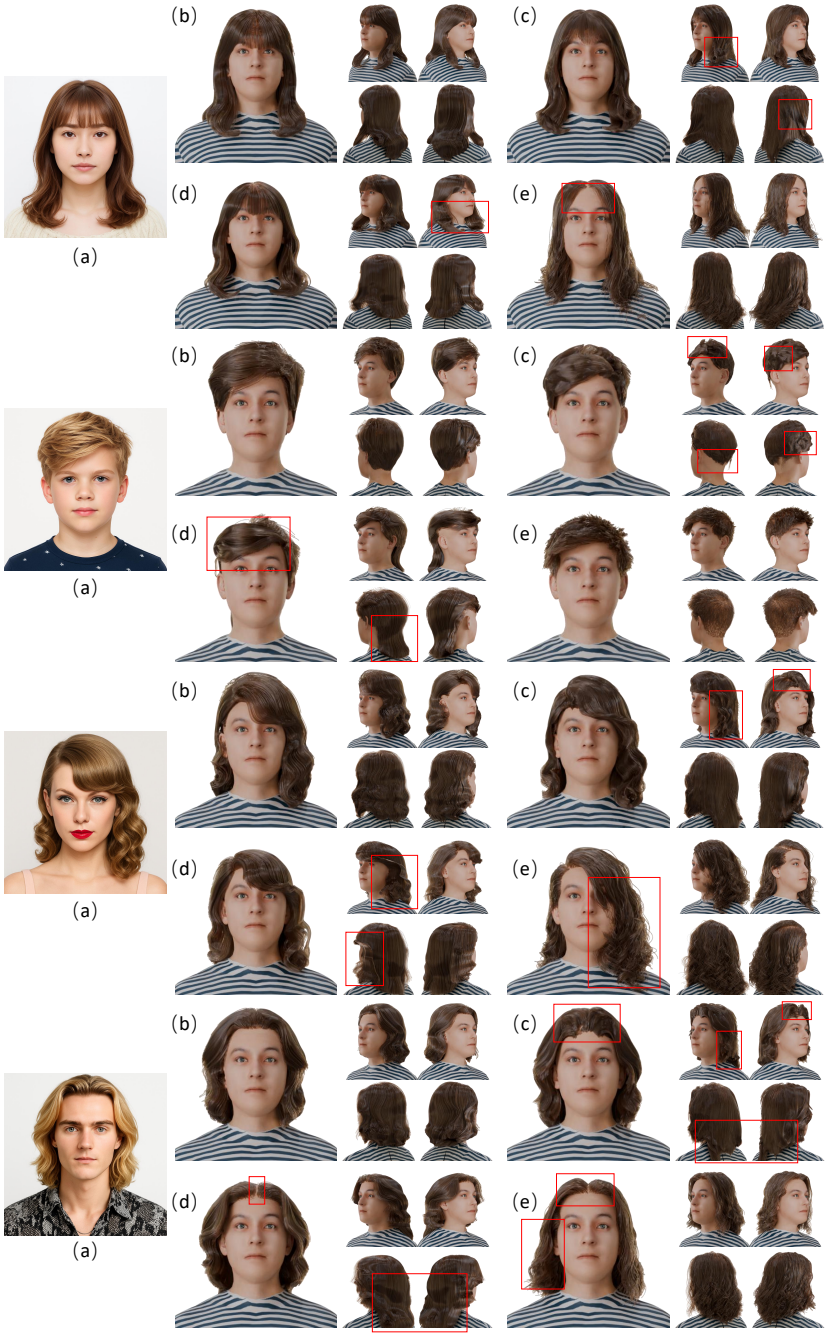

Fig. 5: Qualitative comparison on single-view 3D strands reconstruction. For every example, we show (a) the input image and the reconstructed 3D hair strands rendered in multiple views of (b) Ours, (c) Im2Haircut, (d) HairStep and (e) Difflocks.

quantitatively, we follow [42] to adopt HairSale, HairRida and IoU as the evaluation metrics to assess the single-view alignment by projecting the reconstructed 3D strands onto the image plane to calculate the angle error, accuracy of relative depth and IoU against manually annotated 196 hairstyles. From the quantitative results shown in Tab. 2, we can see our method outperforms on both metrics significantly.

Visual comparisons are also provided in Fig. 5. Im2Haircut and HairStep demonstrate good overall alignment with the input image, benefiting from a cleaner strand map and differentiable rendering–based post-optimization, respectively. However, both methods fail to reconstruct the invisible regions, such as the side and back views, where the length and curliness of the hair are inconsistent with those in the input front view. Moreover, the relative depth and the results obtained from DepthPro [1], which they rely on, are not sufficiently accurate. Consequently, their reconstructions exhibit messy and unsmooth strand geometries, particularly around the frontal–side transition region. Difflocks generates 3D hair strands that appear to match the hairstyle in the input image at first glance. However, it fails to capture the actual hairstyle characteristics presented in the input view, such as the frontal bangs in the first example and the distinct curls and bends in the third and fourth examples. A potential reason is that Difflocks directly generates a scalp texture latent map from the DINOv2 [20] features of the input image through a diffusion process, without establishing sufficient spatial correspondence with the original image. In contrast, our method generates the most accurate and multi-view plausible 3D hair strands compared with previous SOTA methods.

Table 2: Quantitative comparisons on single-view 3D strands reconstruction with HairSale and IoU.

Metric	**HairStep**	**DiffLocks**	**Im2Haircut**	**Ours**
HairSale (°) ↓	17.38	26.50	16.24	**12.83**
HairRida (%) ↑	77.01	69.67	79.38	**80.52**
IoU ↑	0.639	0.593	0.757	**0.847**

4.4 Ablation

To comprehensively investigate the effectiveness of each design within our pipeline, we conduct an ablation study with the following four settings:

- $\boldsymbol{C_0}$: Replace our full-view orientation extractor with the traditional Gabor filter.
- $\boldsymbol{C_1}$: Modify the strand growing method to purely grow strands from the scalp.

– $\boldsymbol{C_2}$: Reduce the conditional views used for 3D field inference to a single input view only (i.e., disables our multi-view bridging strategy).
– ***Full***: Complete *HairOrbit* pipeline.

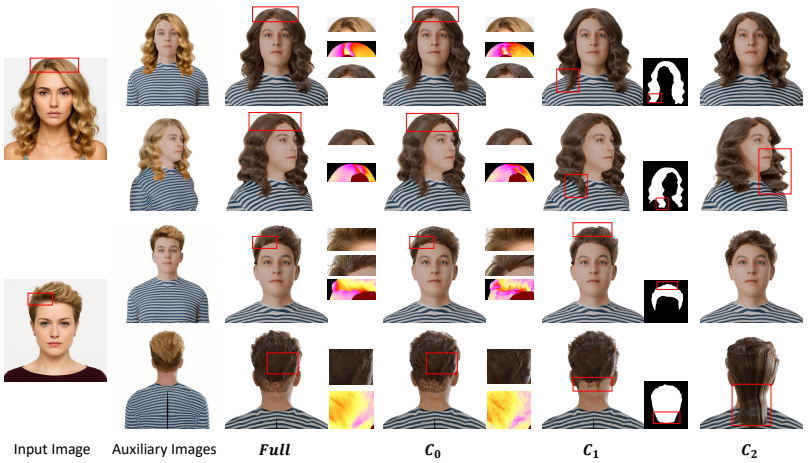

Fig. 6: Qualitative comparisons of ablation study.

Table 3: Quantitative comparisons of the ablation study. Blue values denote the performance degradation (%) compared to ***Full***.

Method	@Input View		@Synthesized 8 Views	
	HairSale (°) ↓	IoU ↑	HairSale (°) ↓	IoU ↑
$\boldsymbol{C_0}$	16.42 (-27.98%)	0.794	13.45 (-34.90%)	0.878
$\boldsymbol{C_1}$	13.70 (-6.78%)	0.828	10.37 (-4.01%)	0.887
$\boldsymbol{C_2}$	19.10 (-48.87%)	0.680	24.14 (-142.1%)	0.658
Full	**12.83**	**0.847**	**9.97**	**0.900**

For each configuration, we report the quantitative results of HairSale and IoU for both the input view and the average across 8 synthetic novel views in Tab. 3. We also provide visual illustration in Fig. 6.

Ablation on full-view orientation extractor Compared with $\boldsymbol{C_0}$, our proposed orientation extractor significantly boosts the reconstruction accuracy on the input view and on the synthesized multiple views (Tab. 3). This substantial improvement indicates that our orientation extractor infers more accurate and view-consistent hair orientations than the traditional Gabor filter. The visual

comparisons between ***Full*** and $\boldsymbol{C_0}$ shown in Fig. 6 further validate these observations. The orientation maps derived from image filters often contain severe noise, which particularly affects key features such as hair partitioning during reconstruction.

Ablation on the strand growing algorithm Compared with $\boldsymbol{C_1}$, our strand growing strategy achieves a better performance (Tab. 3). The modest improvement indicates that scalp-rooted growth already covers the majority of the hair volume. The visual results in Fig. 6 demonstrate that our method reconstructs more accurate strand geometry, particularly yielding more complete structures in the hair boundary and buzz-cut regions.

Ablation of multi-view bridging strategy Compared with $\boldsymbol{C_2}$, the incorporation of our multi-view bridging strategy yields a significant improvement for both the input view and the synthesized multiple views (Tab. 3). The visual comparisons between ***Full*** and $\boldsymbol{C_2}$ in Fig. 6 clearly demonstrate the effectiveness of integrating multi-view constraints in the 3D field inference pipeline, enhancing both local detail and view consistency. Without synthesizing multiple views, $\boldsymbol{C_2}$ cannot estimate an accurate depth map for the input view. Therefore, we use relative depth following [42].

5 Conclusion

In this work, we revisit the long-standing challenge of reconstructing strand-level 3D hair geometry from a single image and reformulate it as a calibrated multi-view reconstruction problem. By leveraging the intrinsic cross-view priors encoded in video diffusion models, our approach synthesizes geometrically consistent and texture-rich multi-view observations from a single portrait. We further propose a full-view neural orientation extractor, a hybrid implicit field representation, and an efficient strand-growing strategy that jointly enhance reconstruction fidelity, completeness, and efficiency. Extensive experiments demonstrate the superiority of our method. Limitations and future works are discussed in the supplementary.

Acknowledgements

This work was supported by the Metaverse Center Grant from the MBZUAI Research Office.

References

1. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024)
2. Chai, M., Shao, T., Wu, H., Weng, Y., Zhou, K.: Autohair: Fully automatic hair modeling from a single image. ACM Transactions on Graphics **35**(4) (2016)
3. Chai, M., Wang, L., Weng, Y., Jin, X., Zhou, K.: Dynamic hair manipulation in images and videos. ACM Transactions on Graphics (TOG) **32**(4), 1–8 (2013)
4. Chai, M., Wang, L., Weng, Y., Yu, Y., Guo, B., Zhou, K.: Single-view hair modeling for portrait manipulation. ACM Transactions on Graphics (TOG) **31**(4), 1–8 (2012)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
6. Hu, L., Ma, C., Luo, L., Li, H.: Robust hair capture using simulated examples. ACM Transactions on Graphics (TOG) **33**(4), 1–10 (2014)
7. Hu, L., Ma, C., Luo, L., Li, H.: Single-view hair modeling using a hairstyle database. ACM Transactions on Graphics (ToG) **34**(4), 1–9 (2015)
8. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
9. Kuang, Z., Chen, Y., Fu, H., Zhou, K., Zheng, Y.: Deepmvshair: Deep hair modeling from sparse views. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022)
10. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
11. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), `https://arxiv.org/abs/2506.15742`
12. Liang, S., Huang, X., Meng, X., Chen, K., Shapiro, L.G., Kemelmacher-Shlizerman, I.: Video to fully automatic 3d hair model. ACM Transactions on Graphics (TOG) **37**(6), 1–14 (2018)
13. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36**, 22226–22246 (2023)
14. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
15. Luo, L., Li, H., Paris, S., Weise, T., Pauly, M., Rusinkiewicz, S.: Multi-view hair capture using orientation fields. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1490–1497. IEEE (2012)
16. Luo, L., Zhang, C., Zhang, Z., Rusinkiewicz, S.: Wide-baseline hair capture using strand-based refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 265–272 (2013)
17. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
18. Nam, G., Wu, C., Kim, M.H., Sheikh, Y.: Strand-accurate multi-view hair capture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 155–164 (2019)

19. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Paris, S., Briceno, H.M., Sillion, F.X.: Capture of hair geometry from multiple images. ACM transactions on graphics (TOG) **23**(3), 712–719 (2004)
22. Paris, S., Chang, W., Kozhushnyan, O.I., Jarosz, W., Matusik, W., Zwicker, M., Durand, F.: Hair photobooth: geometric and photometric acquisition of real hairstyles. ACM Trans. Graph. **27**(3), 30 (2008)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Rosu, R.A., Wu, K., Feng, Y., Zheng, Y., Black, M.J.: Difflocks: Generating 3d hair from a single image using diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10847–10857 (2025)
26. Saito, S., Hu, L., Ma, C., Ibayashi, H., Luo, L., Li, H.: 3d hair synthesis using volumetric variational autoencoders. ACM Transactions on Graphics (TOG) **37**(6), 1–12 (2018)
27. Shen, Y., Zhang, C., Fu, H., Zhou, K., Zheng, Y.: Deepsketchhair: Deep sketch-based 3d hair modeling. IEEE transactions on visualization and computer graphics **27**(7), 3250–3263 (2020)
28. Sklyarova, V., Chelishev, J., Dogaru, A., Medvedev, I., Lempitsky, V., Zakharov, E.: Neural haircut: Prior-guided strand-based hair reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19762–19773 (2023)
29. Sklyarova, V., Zakharov, E., Hilliges, O., Black, M.J., Thies, J.: Text-conditioned generative model of 3d strand-based human hairstyles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4703–4712 (2024)
30. Sklyarova, V., Zakharov, E., Prinzler, M., Becherini, G., Black, M.J., Thies, J.: Im2haircut: Single-view strand-based hair reconstruction for human avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10656–10665 (2025)
31. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
32. Tang, Z., Geng, J., Weng, Y., Zheng, Y., Zhou, K.: Single-view 3d hair modeling with clumping optimization. IEEE Transactions on Visualization and Computer Graphics (2025)
33. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
34. Wu, K., Yang, L., Kuang, Z., Feng, Y., Han, X., Shen, Y., Fu, H., Zhou, K., Zheng, Y.: Monohair: High-fidelity hair modeling from a monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24164–24173 (2024)

35. Wu, K., Ye, Y., Yang, L., Fu, H., Zhou, K., Zheng, Y.: Neuralhdhair: Automatic high-fidelity hair modeling from a single image using implicit neural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1526–1535 (2022)
36. Yang, L., Shi, Z., Zheng, Y., Zhou, K.: Dynamic hair modeling from monocular videos using deep neural networks. ACM Transactions on Graphics (TOG) **38**(6), 1–12 (2019)
37. Zakharov, E., Sklyarova, V., Black, M., Nam, G., Thies, J., Hilliges, O.: Human hair reconstruction with strand-aligned 3d gaussians. In: European Conference on Computer Vision. pp. 409–425. Springer (2024)
38. Zhang, M., Chai, M., Wu, H., Yang, H., Zhou, K.: A data-driven approach to four-view image-based hair modeling. ACM Trans. Graph. **36**(4), 156–1 (2017)
39. Zhang, M., Wu, P., Wu, H., Weng, Y., Zheng, Y., Zhou, K.: Modeling hair from an rgb-d camera. ACM Transactions on Graphics (TOG) **37**(6), 1–10 (2018)
40. Zhang, M., Zheng, Y.: Hair-GAN: Recovering 3D hair structure from a single image using generative adversarial networks. Visual Informatics **3**(2), 102–112 (2019)
41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
42. Zheng, Y., Jin, Z., Li, M., Huang, H., Ma, C., Cui, S., Han, X.: Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12726–12735 (2023)
43. Zheng, Y., Qiu, Y., Jin, L., Ma, C., Huang, H., Zhang, D., Wan, P., Han, X.: Towards unified 3d hair reconstruction from single-view portraits. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
44. Zhou, Y., Hu, L., Xing, J., Chen, W., Kung, H.W., Tong, X., Li, H.: Hairnet: Single-view hair reconstruction using convolutional neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 235–251 (2018)
45. Zhou, Y., Chai, M., Wang, D., Winberg, S., Wood, E., Sarkar, K., Gross, M., Beeler, T.: Groomcap: High-fidelity prior-free hair capture. ACM Transactions on Graphics (TOG) **43**(6), 1–15 (2024)