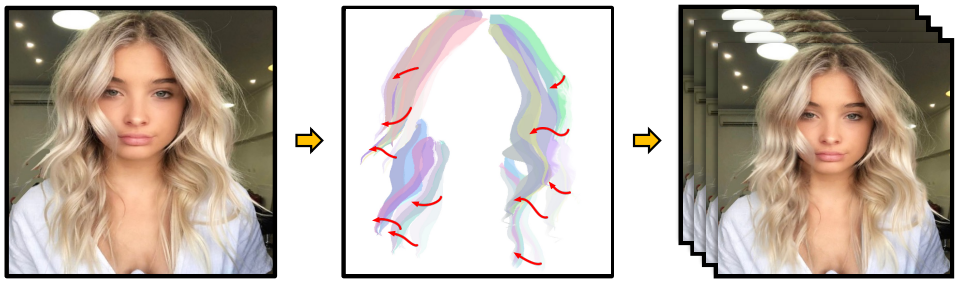
Automatic Animation of Hair Blowing in Still Portrait Photos

Wenpeng Xiao[1], Wentao Liu[1], Yitong Wang[1], Bernard Ghanem[2], Bing Li[2]

[1] Bytedance Inc. [2] King Abdullah University of Science and Technology

{xiaowenpeng.com, liuwentao.canon}@bytedance.com, wangyitong@pku.edu.cn, {Bernard.Ghanem, bing.li}@kaust.edu.sa



(a) (b) (c)

Figure 1: Given a still portrait photo (a), our method automatically detects hair wisps and animates wisps (b), while converting the photo into a cinemagraph (c).

Abstract

We propose a novel approach to animate human hair in a still portrait photo. Existing work has largely studied the animation of fluid elements such as water and fire. However, hair animation for a real image remains underexplored, which is a challenging problem, due to the high complexity of hair structure and dynamics. Considering the complexity of hair structure, we innovatively treat hair wisp extraction as an instance segmentation problem, where a hair wisp is referred to as an instance. With advanced instance segmentation networks, our method extracts meaningful and natural hair wisps. Furthermore, we propose a wisp-aware animation module that animates hair wisps with pleasing motions without noticeable artifacts. The extensive experiments show the superiority of our method. Our method provides the most pleasing and compelling viewing experience in the qualitative experiments, and outperforms state-of-the-art still-image animation methods by a large margin in the quantitative evaluation. Project url: https://nevergiveu.github.io/AutomaticHairBlowing/

1. Introduction

"His silvery hair was blowing in the wind," George R.R. Martin[1] described. Hair is one of the most impressive parts of the human body [61, 77], while its dynamics make a deeper impression and make the scene vivid. The studies show that dynamics is more compelling and captivating than a still image. Massive portrait photos are shared every day on social media platforms such as TikTok and Instagram. People want their photos to be attractive and artistic. This motivates us to explore animating human hair in a still image, so as to provide a vivid, pleasing and beautiful viewing experience. Recent methods [19, 52, 26] have been proposed to augment a still image with dynamics, which animates fluid elements such as water, smoke and fire in the image. However, these methods haven't taken human hair into account for real photos.

To provide an artistic effect, we focus on animating human hair in a portrait photo, while translating the photo into a ***cinemagraph*** [68, 1]. Cinemagraph is an innovative short-video format preferred by professional photographers, advertisers, and artists, and it is used in digital advertisements, social media, landing pages, etc. The engaging nature of the cinemagraph is that it integrates the merits of still photos and videos [56, 52, 4]. That is, some regions in a cinemagraph contain small motions in a short loop, while the rest remain static. The contrast between static and moving elements helps to capture viewers' attention. Translating a portrait photo into a cinemagraph with subtle hair motions would make the photo to be more compelling yet not distract viewers from static content.

[1] Celebrated novelist of "Wild Cards", "A Game of Thrones", etc. [75]

Existing methods (*e.g*. [56, 84]) and commercial software ([20, 22, 2]) generate a high-fidelity cinemagraph from an input video by freezing selective video regions. These tools are not applicable to a still image. On the other hand, still-image animation has attracted increasing attention [19, 52, 26]. Most approaches explore animating fluid elements such as clouds, water, and smoke. However, hair is made of fibrous materials, leading to its dynamics being rather different from that of fluid elements. Different from fluid element animation which has been largely investigated, human hair animation is much less been explored for a real portrait photo.

Animating hair in a still portrait photo is a challenging problem. Research on hair modeling, which aims to reconstruct plausible hair for virtual humans, has revealed the high complexity of hair in terms of structure and dynamics. For example, different from human body or face that has smooth surfaces, hair comprises around hundred-thousand components *i.e*. *hair strands*, leading to intricate structures. Furthermore, such a massive number of fibrous components result in non-uniform and complicated motions within the hair as well as temporal collisions between hair and head. Many hair modeling approaches address hair complexity by resorting to specialized hair capture techniques (*e.g*. dense camera array and high-speed cameras). Thus, static hair modeling approaches [30, 51, 54, 62] construct high-quality 3D models of static hair, and dynamic hair modeling approaches [72, 83] achieve impressive results in reconstructing hair motions at strand level. However, these approaches suffer from expensive time costs or rely on complex hardware setups to capture real-world hair.

In this paper, we propose a novel method that automatically animates hair in a still portrait photo without any user assistance or sophisticated hardware. We observe that human visual system is much less sensitive to hair strands and their motions in a real portrait video, compared to synthetic strands of a digitized human in a virtual environment. Our insight is that we can animate hair wisps rather than strands which can create a perceptually pleasing viewing experience. We hence propose a hair wisp animation module to animate hair wisps, enabling an efficient solution.

The arising challenge is how to extract hair wisps. Although relevant work such as hair modeling investigates hair segmentation, these approaches focus on extracting the whole hair region, which is different from our aim. To extract meaningful hair wisps, we innovatively treat hair wisp extraction as an instance segmentation problem, where a segment instance from a still image is referred to as a hair wisp. Thanks to such a problem definition, we can exploit instance segmentation networks to realize the extraction of hair wisps. This largely simplifies the hair wisp extraction problem, but also advanced networks can effectively extract hair wisps. Furthermore, we construct a hair wisp dataset that consists of real portrait photos to train the networks. We also propose a semi-annotation scheme to produce ground-truth annotations of hair wisps.

Our contributions are summarized as follows:

- We propose a novel approach that automatically animates blowing hair from a still portrait image. Our method effectively handles high-resolution images, while generating high-quality and aesthetically-pleasing cinemagraphs without any user assistance.
- We show that instance segmentation facilitates the animation of hair blowing, which is helpful in generating realistic blowing motions.
- We propose a hair wisp animation module that generates pleasing motions for hair wisps without noticeable artifacts.

2. Related Work

Hair Modeling. Many efforts have been devoted to hair modeling which is to generate/reconstruct human hairs for virtual humans. Hair modeling from scratch is laborious and time-consuming. Modeling methods [14, 38, 87] are proposed to synthesize hair strands to ease manual work. Detailed discussions can be found in the survey[73]. To narrow down the gap between synthetic and real hair, hair capture methods are introduced to construct hair models from the real world. These methods can be roughly categorized into static hair capture and dynamic one. Static hair capture methods model static hair using various sensors. Methods [57, 35, 5, 51, 54, 62, 57, 39] deploy 3D hair models from multi-view images or point clouds. However, most hair capture methods require 3D acquisition devices while relying on complex hardware setups. Instead, a few work has explored reconstructing hair from a single-view image [9, 11, 31, 11, 93, 10, 89, 78] or sparse views [39].

A few work [90, 83, 85, 72, 29, 17] has explored dynamic hair capture. Most methods [90, 83, 72, 29] employ specialized hardware techniques such as calibrated camera array and lighting control to capture hair dynamics, where Wang *et al*. [72] propose a hair-tracking algorithm using multi-view videos. Zhang *et al*. [90] and Hu *et al*. [29] exploit physical simulation for hair modeling. Winberg *et al*. [76] employ 14 synchronized cameras to track 3D facial hair and the underlying 3D skin. Yang *et al*. [85] introduce the first deep-learning networks for dynamic hair modeling, where synthetic hair data are used to train the networks. Different from these hair modeling approaches designed for virtual humans, our work focuses on automatically animating hair for a real portrait photo.

Single-image-to-video generation. Different video-based synthesis methods (*e.g*. [4, 49, 56, 68, 86]) generating video from the input video, many methods [19, 52,

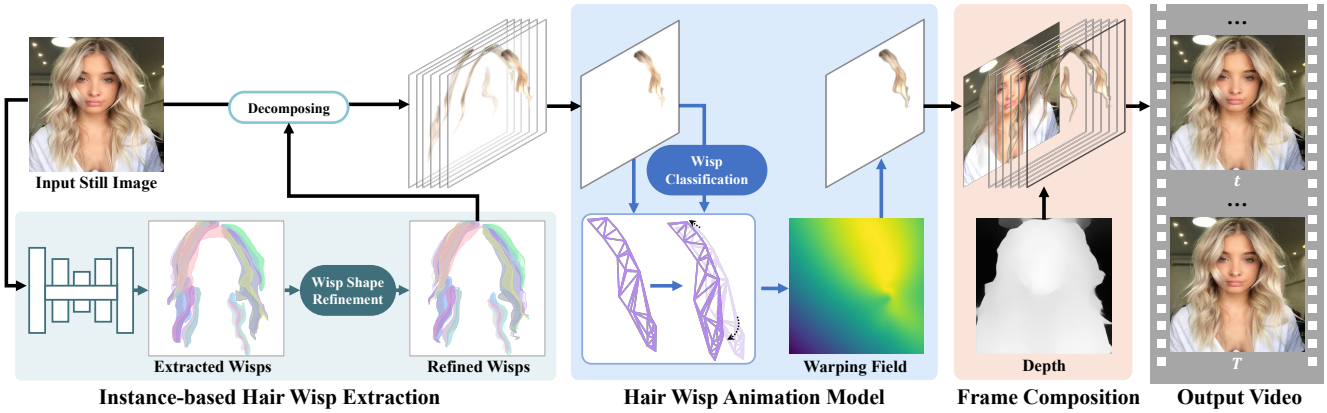


Figure 2: The framework of our method. Given a still photo, our instance-based hair wisp extraction module extracts hair wisps from it. The proposed hair wisp animation module then represents the extracted hair wisps by multi-layer mesh representation and creates motions for each extracted wisp by driving the wisp's mesh. Animated hair wisps and the background are composited to render frame $\tilde{I}^t$ for generating a video using the depth information of the photo.

15, 26, 21, 3, 6, 8, 18, 74, 27, 55] have been proposed to convert a still image into a video. As generative models have shown impressive performance in image and video translation, video prediction methods [6, 8, 18, 32] predict video frames from a single one using Generative Adversarial Networks (GANs), variational autoencoders, conditional invertible neural network. These methods often suffer low-resolution issues, since understanding and generating videos are challenging [82, 91]. Differently, methods [15, 24] animate a still image by estimating motion fields. With manually segmented layers, Chuang et al. [15] animate a layer with a stochastic motion texture using harmonic oscillations [64] Hao *et al*. [24] generate a video from an image while taking additional sparse motion trajectories as input, where motion trajectories are provided by users. Halperin *et al*. combine conditional random field with a local feature descriptor to calculate dense displacement fields from a user-provided motion direction. Mahapatra and Kulkarni [52] generate a dense optical flow map from user guidance by exponential functions and GANs. These methods require user assistance such as motion directions or a manual mask indicating which regions need to be animated. Recent work [19, 26] design motion estimation networks to predict motion field. Holynski *et al*. [26] focus on animating fluid elements, where the motions of a generated video are represented by a static motion field. Methods [21, 3, 45, 81] have devoted efforts to portrait image animation, where some methods [79, 81] utilize attribute-level information [48, 47]. However, these approaches mainly focus on editing facial images, but also rely on additional reference videos or 3D face models. Different from these approaches, our method focuses on animating human hair in a still image without user assistance.

3. Methodology

Problem Definition. Given a still portrait photo I, our aim is to automatically generate a cinemagraph $V = \{\tilde{I}^t\}_{t=1}^{T}$, while animating human hair in the photo, where T is frame number and $\tilde{I}^t$ is a frame of the generated video. However, animating hair in a still image is challenging, since the complexity of hair structure and dynamics poses new challenges compared with fluid element animation.

We address the above challenges by exploring two questions: (1) What to animate in human hair? (2) How to automatically and naturally "blow" hair? Towards the first question, we observe that human visual system is much less sensitive to hair-strand-level motion than the wisp-level one in real short videos, different from virtual/digitized environments. Motivated by this, we propose a hair wisp extraction module to extract meaningful hair wisps for still portrait photos and a wisp animation module that animates extracted hair wisps with natural and pleasing motions.

Overview. We propose a framework for animating hair for still portrait images. Our framework consists of three steps, as shown in Fig. 2. First, we propose an Instance-based Hair Wisp Extraction (IHWE) which automatically extracts hair units named *hair wisps* that are locally grouped and would move consistently in a generated video, without relying on complex hair capture systems or user assistance. Second, we propose a hair wisp animation module to animate hair wisps by predicting the spatiotemporal evolution of a hair wisp. Third, with the animated hair wisps, we generate an animated video by fusing animated hair wisps.

3.1. Instance-based Hair Wisp Extraction

We propose to automatically extract hair wisps from a portrait photo for animating hair. To the best of our knowl-

edge, although some hair capture approaches [9, 88] and face parsing approaches [50, 67, 92, 53, 46] have deployed hair segmentation algorithms for a single image, these approaches mainly focus on extracting the whole hair region, rather than hair wisps. Due to the intricate appearance and structure of hair, it is nontrivial to accurately segment the whole hair region [9], while automatically extracting hair wisps is much more challenging. For example, hair wisps are visually similar to each other in the same hair (see Fig. 1), which needs extraction methods to discriminate subtle differences among them.

Different from existing work, we cast hair wisp extraction as an instance segmentation problem, inspired by the remarkable performance of supervised instance segmentation methods on animals and humans. Thus, by treating a wisp as an instance, we can employ advanced instance segmentation networks to extract hair wisps. Nevertheless, the difficulties lie in training data, especially the ground-truth annotations. It is time-consuming and expensive to annotate each hair wisp in a real image manually. To address this challenge, we propose a training data construction scheme. Below, we elaborate on the proposed instance-based wisp extraction module and data construction scheme.

Hair Wisp Extraction. Given a still portrait image I, we present an IHWE module to predict instance masks $\mathbf{M} = \{M_i\}_{i=1}^{N}$ for hair wisps, where N is the instance number, $M_i \in R^{W \times H}$ is an instance mask of a hair wisp, $W \times H$ is the size of image I. In particular, we first predict a matting map $\bar{M}$ which indicates the whole hair regions using [13], to avoid irrelevant components negatively affecting instance segmentation results. With the matting map, we employ deep neural networks [28] to extract hair wisps, inspired by the success of supervised instance segmentation methods (*e.g.*, [28] [34]). Thanks to our instance segmentation, hair wisps are adaptively extracted according to hair content, without pre-defining wisp number N. After that, we further refine the shape of extracted hair wisps.

The issue is that these networks are task-oriented and rely on specific training data. However, there are no proper hair wisp datasets for instance segmentation. Although a few hair datasets ([37, 80]) have been presented, these datasets are constructed for extracting the whole hair region (*e.g.* face parsing), which is not applicable to our task.

Hair Wisp Dataset. We constructed a new hair dataset named Hair-Wisp dataset for instance segmentation, such that instance segmentation networks can be trained to detect hair wisps from portrait images in a supervised manner. However, it is laborious to manually annotate a segmentation mask for each wisp. To address this issue, we first propose a sketch-filled algorithm to generate ground-truth annotations of hair wisps for a portrait image.

Our sketch-filled algorithm generates ground-truth annotations of hair wisps inspired by *guide* hairs used in hair modeling. In particular, existing hair modeling methods [58, 12] generate/synthesize hair from a small number of guide hair strands that are representative to depict hairstyles/structures. Inspired by this, we found a hair sketch dataset [80] which provides user strokes that well indicate the distribution of hair wisps. Moreover, xiao *et al*. [80] synthesizes plausible hair using hair sketch. Here, given a portrait image in hair sketch dataset, we exploit the associated hair sketch as a guided hair strand and annotate hair wisp by the sketch.

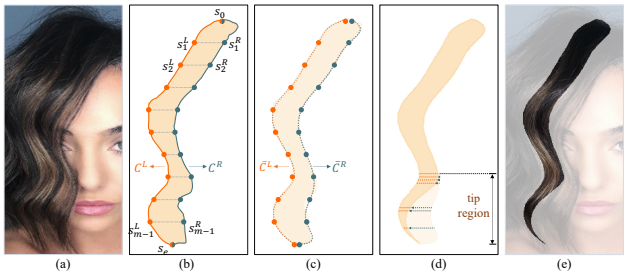


Figure 3: Illustration of hair wisp shape refinement. Given a hair wisp in (b) extracted from an image (a) via instance segmentation, our method smooths the left contour C^L and the right one C^L in (b), and obtains refined left/right contours ($\bar{C}^L$, $\bar{C}^R$) in (c). We then sharpen the tip region of the wisp in (d).

One way of generating wisp annotations is to directly expand the hair sketch. However, the generated annotation results in such a manner are low-quality. For example, the annotation results improperly annotate parts of two neighboring wisps as a wisp. In addition, the annotated regions of hair wisps are rugged and lack smoothness. Instead, we design a sketch-filled algorithm based on flood-fill [42] by expanding in the top, bottom, and right directions. By ignoring the left-direction expansion, our algorithm well preserves the contour indicated by the strokes.

With the above annotation generation, we construct the Hair-Wisp dataset from the hair sketch dataset [80]. Our dataset includes 4500 images and ground-truth annotations of hair wisps, which covers challenging cases such as braid hairstyles. It is worth noting that our method does need completely accurate annotations to train hair wisp extraction networks, as we design shape refinement algorithms below. Please refer to the supplementary for more details.

Hair Wisp Shape Refinement. After exploiting instance segmentation networks to extract hair wisps from the input photo, we further refine the extracted hair wisps. Due to the limitation of instance segmentation networks, the contours of an extracted hair wisp is not smooth and its tip is coarse. (see Fig. 3 (a)). Hence, we smooth the contours of extracted hair wisps and then sharpen wisp tips. Given an extracted wisp, we first split its contour into left and right

contours using its top-most and bottom-most points. The left/right contour of extracted hair is uniformly divided to obtain sample points (see Fig. 3 (b)). Let $\{s_0, s_1^z, ..., s_e\}$ be sample points on z contour C^z, where $z = L$ denote left and $z = R$ right. The left/right contour is smoothed by polynomial regression g:

$$\bar{C}^z = g(s_0, s_1^z, ..., s_e) \tag{1}$$

We then sharpen the tip region of a hair wisp by linearly shrinking the width of the tip region (see Fig. 3 (d)).

3.2. Hair Wisp Animation

We describe how to generate motions that animate human hair wisps in a still image. Recent single-image-to-video methods [52, 26] use deep learning techniques to animate fluid elements, since fluid motion can be approximated via a static velocity field. However, since hair motions are complex [71] and the input is only a single image, it is non-trivial to create hair motions using deep learning techniques. Instead, we animate hair wisps based on physical models.

There are two new challenges posed by generating hair wisp motions. First, the dynamics of a hair wisp not only include motion displacement but also the shape deformation of the wisp. As a result, if we apply physically based animation algorithms (*e.g.* simulation [63, 17]) to hair wisps by treating a hair wisp as a strand, the algorithm fails to model shape deformation, leading to unrealistic motions and artifacts. Second, the hair root regions of many extracted hair wisps are partially occluded by other objects such as the face in a portrait photo. However, the motion of a hair wisp is continuous and starts at the scalp of the head. In other words, given a wisp, its local region closer to the head scalp affects the motions of other farther regions in it. Hence, it is difficult to create motions for partially-occluded hair wisps.

We address these challenges by proposing a hair wisp animation module. In particular, to model shape deformation and motion displacement, we represent hair wisps with meshes. We then impose mass-spring systems on meshes to temporally drive the hair wisps for animation, since mass-spring systems have been used to simulate various non-rigid objects such as cloth, providing a simple yet effective solution. Furthermore, we recognize partially-occluded hair wisps that are not connected to the head scalp, and explicitly approximate motions for these wisps.

Wisp classification. We classify hair wisps into two categories: scalp-connected and scalp-unconnected. In particular, we first obtain the forehead contour C of the human face in input image I by detecting facial keypoints along the forehead contour and connecting them. If a hair wisp intersects with the forehead contour, it is classified as scalp-wisp; otherwise non-scalp-wisp.

Multi-layer Mesh Representation. Given an input image I, we represent hair wisps in I by multiple layers of meshes, rather than a single mesh. In particular, the k-th hair wisp W^k is represented by a triangle mesh $\{\mathbf{X}^k, \mathbf{E}^k\}$, where $\mathbf{E}^k$ is the set of triangle edges, $\mathbf{X}^k = \{\mathbf{x}_i^k\}$ is the set of mesh vertexes and $\mathbf{x}_i^k \in \mathbb{R}^2$ is the postion of the i-th vertex. We construct a mesh by Delaunay Triangulation. To improve the freedom of degree (FOD) of a mesh, we increase the number of vertexes in a mesh by dividing circumscribed rectangles of the hair wisps into 6×6 grids.

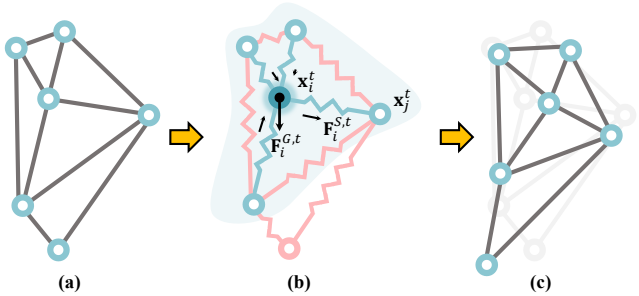


Figure 4: Given a mesh of a hair wisp (a), springs that are imposed to connect neighboring vertexes $\mathbf{x}^t$ in (b) drive the mesh by forces F, such that warped mesh (c) approximates the spatiotemporal evolution of the wisp.

Based on the above mesh representation, predicting the dynamics of a hair wisp W^k is to predict the spatial-temporal evolution of its mesh along the time axis. Let $\tilde{\mathbf{X}}^{k,t}$ be a set including positions of vertexes $\mathbf{X}^k$ at time t. After obtaining $\tilde{\mathbf{X}}^{k,t}$, we can generate the animated hair wisp $\tilde{W}^{k,t}$ for a generated frame $\tilde{I}^t$ by warping W^k:

$$\tilde{W}^{k,t} = \varpi(W^k, \mathbf{X}, \tilde{\mathbf{X}}^{k,t}), \tag{2}$$

where ϖ is the warping operator. Here, we employ the thin plate spline algorithm [7] to obtain dense warping fields.

Wisp Motion Prediction. To estimate vertex positions $\tilde{\mathbf{X}}^{k,t}$ at time t and simulate motions for hair wisp W^k, we impose mass-spring systems on the mesh of W^k . In particular, we directly use a mesh vertex as a particle with mass m, and each two neighboring particles are connected by a spring (see Fig. 4). Consequently, the motion of a vertex is affected by spring forces and other forces. Here, we only consider gravity and spring forces for the sake of simplicity. Below, We omit the superscript k for brevity. The accumulated force $\mathbf{F}^t$ that acts on vertex $x_i^t \in \tilde{\mathbf{X}}^{k,t}$ is measured as the gravity force $\mathbf{F}_i^{G,t}$ and spring force $\mathbf{F}_i^{S,t}$:

$$\mathbf{F}_i^t = \mathbf{F}_i^{G,t} + \mathbf{F}_i^{S,t} \tag{3}$$

The spring force $\mathbf{F}_i^{S,t}$ is computed according to Hooke's law. Since a vertex is connected by multiple springs, we compute $\mathbf{F}_i^{S,t}$ from all springs connected to $\mathbf{x}_i^t$:

$$\mathbf{F}_i^{S,t} = \sum_{j \in \mathcal{N}} K \cdot (\mathbf{x}_i^t - \mathbf{x}_j^t - |\mathbf{e}_{ij}|) \tag{4}$$

where K is the spring constant indicating force strength, $\mathbf{e}_{ij}$ is a spring connected to the i-th vertex $\mathbf{x}_i^t$, $|\mathbf{e}_{ij}|$ is the spring length, $\mathcal{N}$ is the set including the indexes of neighboring vertexes of $\mathbf{x}_i^t$.

With the accumulated force, the vertex would be moving. We thereby predict vertex positions at time t by estimating acceleration and velocity. In particular, according to Newton's second law of motion, the acceleration of a vertex at time t is calculated:

$$a^t = \frac{\mathrm{d}\mathbf{v}_i(t)}{\mathrm{d}t} = \frac{\mathbf{F}_i^t}{m} \tag{5}$$

where v_i^t is the velocity of vertex $\mathbf{x}_i^t$. The position of the mass meets ODE: $\mathbf{v}_i(t) = \frac{\mathrm{d}\mathbf{x}_i(t)}{\mathrm{d}t}$. By using Euler's method in Eq. 5, we predict vertex positions at the time axis as:

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + a_i^t \cdot \Delta t \tag{6}$$

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x}_i(t) + \mathbf{v}_i(t + \Delta t) \cdot \Delta t \tag{7}$$

where initial velocity $\mathbf{v}_i(0)$ is preset. We can set $\mathbf{v}_i(0) = [0\ 0]$ or set $\mathbf{v}_i(0)$ to be a non-zero vector to indicate wind force. In addition, since vertexes on the head scalp are static, the positions of these vertexes are fixed and are equal to the original values.

We can not directly apply the above wisp motion model to unconnected-scalp wisps. Since no vertex of an unconnected-scalp wisp is fixed, gravity would pull down all vertexes in the wisp, forcing the wisp to keep falling. We address the issue by approximating motions for the topmost vertex of an unconnected-scalp wisp. In particular, we construct a mesh on the entire hair, where the mesh is named auxiliary mesh. We then use the wisp motion model to predict the motion of the auxiliary mesh, while fixing the positions of vertexes lying on the head scalp. For the topmost vertex of an unconnected-scalp wisp, we approximate its motion by that of the auxiliary mesh's vertex which has the same position as it. By such approximation, the topmost vertex of an unconnected-scalp wisp is relatively fixed to the auxiliary mesh. We apply the wisp motion model to the remaining vertexes of an unconnected-scalp wisp.

Note that existing physically-based hair animation methods (*e.g.* [10]) are mainly designed for reconstructed hair stands or virtual humans. These methods first simulate motions for a small number of guide hair strands, and then approximate motions for the rest strands from that of the guide ones. However, it is extremely difficult to accurately extract guide strands without user assistance (*e.g.* [10]), given only a still image. Differently, our method generates motions for hair wisps and automatically detects hair wisps, which largely reduces the computational difficulty and is simpler to apply without user assistance.

3.3. Video Generation

We describe the progress of generating a video from the extracted hair wisps and their predicted dynamics.

Depth-aware frame composition. A naive way of rendering a frame $\tilde{I}^t$ is to composite the background and the warped versions $\{\tilde{W}^{1,t}, ..., \tilde{W}^{k,t}, ...\}$ of all extracted hair wisp, where the background is extracted by removing hair regions from original image I. However, this would lead to improper occlusion relationship between wisps and face. For example, after the composition, some warped hair wisps that should be behind the face may occlude the face, which often leads to unrealistic motions. To address this issue, we introduce depth information to guide the composition [43, 44]. In particular, we use a face parsing algorithm [13] to extract face regions. By applying a depth estimation algorithm [60] to image I, we obtain the relative depth relationship between the face and a hair wisp. The hair wisp layers and background are sorted by the depth and the height of hair wisps. We then composite the sorted warped wisps and the background to render frame $\tilde{I}^t$, following Painter's algorithm. In addition, our method refines the background layer by inpainting [65] missing regions, which avoids hole-filling for each generated frame.

4. Experiments

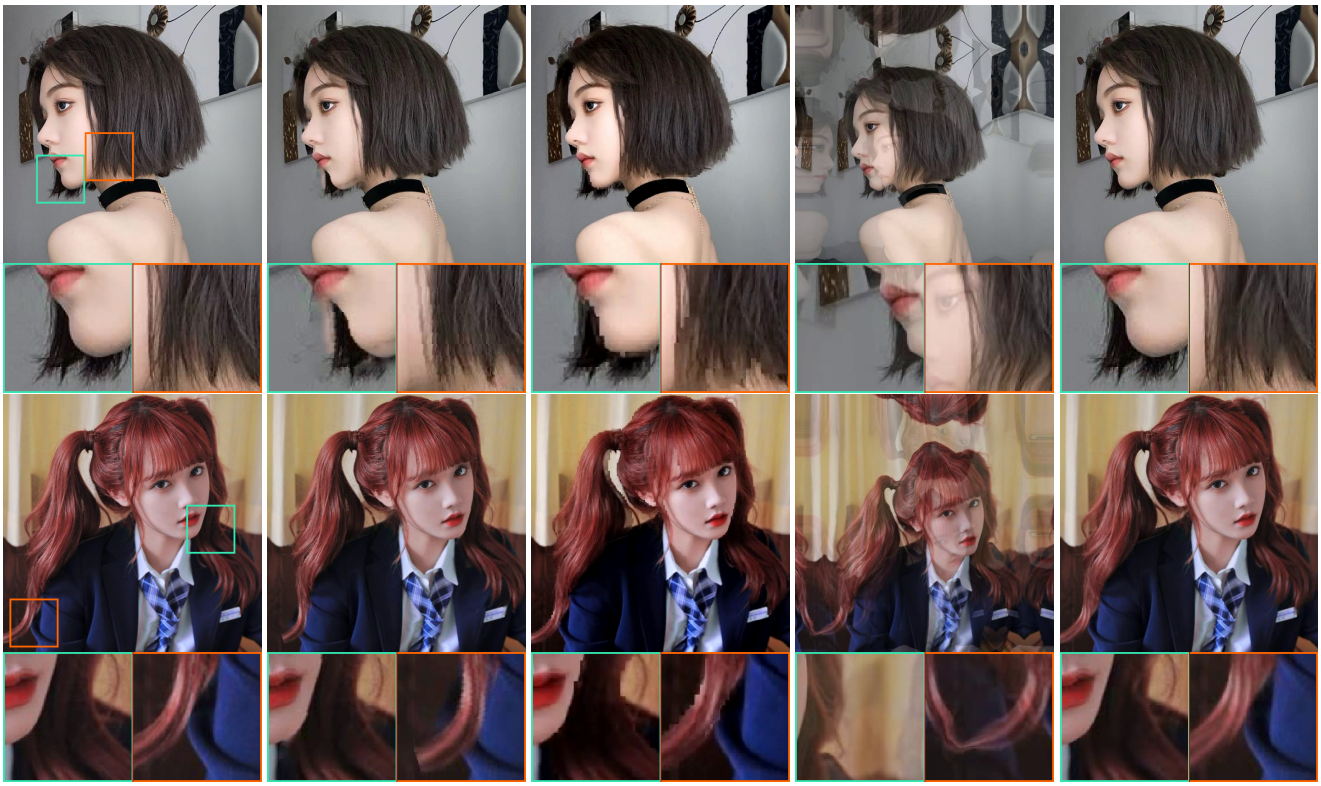
4.1. Baselines and Implementation Details

We compare our method with three state-of-the-art methods in single-image-to-video generation: Halperin *et al.* [23], Endo *et al.* [19] and Chuang *et al.* [15][2], where Halperin *et al.* and Endo *et al.* are recent work using deep learning techniques. Both Endo *et al.* and Chuang *et al.* require user assistance, where Chuang *et al.* needs users to manually decompose the input image into layers, and Halperin *et al.* requires a user-provided mask indicating which regions are moving and a general motion direction. Therefore, we manually specify motion directions and mask hair regions for Chuang *et al.* and Halperin *et al.* in our experiments. To implement Endo *et al.*, we train its networks on real cinemagraphs containing hair blowing. More implementation details and comparisons to other competitors are provided in supplementary materials, where we additionally show the results of applying our method to anime images and clothes.

4.2. Evaluation Metrics

We use two evaluation metrics to assess the performance of our method on video quality and temporal consistency.

[2]To the best of our knowledge, most hair animation methods are designed for synthetic data (*e.g.* games and virtual reality). These methods are not applicable to a real portrait photo, since they require hair strand information which is significantly difficult to acquire without complex hair capture systems.



Input image | Chuang *et al*. [15] | Halperin *et al*. [23] | Endo *et al*. [19] | Ours

Figure 5: Qualitative comparisons of our approach with single-image-to-video generation methods *et al*. [15] [23] [19]. In each image, we show a full frame at the top and zoom-in rectangle regions marked by red/green at the bottom.

Frechet Video Distance (FVD)[69]. FVD is a standard metric that has been popularly used for evaluating the quality of generated videos in recent video synthesis work (*e.g*. [52, 18]). FVD assesses the quality of generated videos by measuring the data distribution between generated and real videos, based on FID [25]. The pre-trained I3D networks are employed [66] to extract video features, where the networks are trained on Kinetics dataset [36]. A lower FVD score indicates a higher quality of generated videos.

Warping Error E_{warp} [40]. We adopt E_{warp} to measure the temporal inconsistency of generated videos, following [41, 16]. To evaluate short- and long-term consistency of a video, E_{warp} computes the warping errors between consecutive frames as well as that of each frame and the first frame. A lower E_{warp} value indicates better temporal consistency.

4.3. Comparison

Quantitative and Qualitative Results. Tab. 1 shows our method achieves the best FVD and E_{warp} value, indicating our method outperforms the three state-of-the-art methods in terms of video quality and temporal consistency. The supplementary video and Figs. 5 and 6 demonstrate qualitative comparison results, showing our approach outperforms state-of-the-art methods on various testing images. More comparisons are provided in supplementary materials.

Endo *et al*. leads to noticeable distortions in Fig. 5 and supplementary video. This is because Endo *et al*. employ networks to learn motion fields from training data, while it is difficult to directly train the networks that can effectively capture hair dynamics due to the complex motion space of hair dynamics. Halperin *et al*. combines a conditional random field with a local feature descriptor to compute a temporally and spatially continuous displacement field, which achieves remarkable animation performance for objects with periodic structures. However, since such a displacement field improperly drives all hair regions to consistently move in the same direction, Halperin *et al*. create unrealistic and unnatural motions for hair (see the supplementary video). In addition, Halperin *et al*. introduces jagged artifacts into the boundary of the face and hair (see Fig. 5). Chuang *et al*. often leads to jitter artifacts in animated hairs, since this method approximates motion as harmonic oscillations, which is more suitable for branches and grass rather than hair. In contrast, our instance-based hair wisp

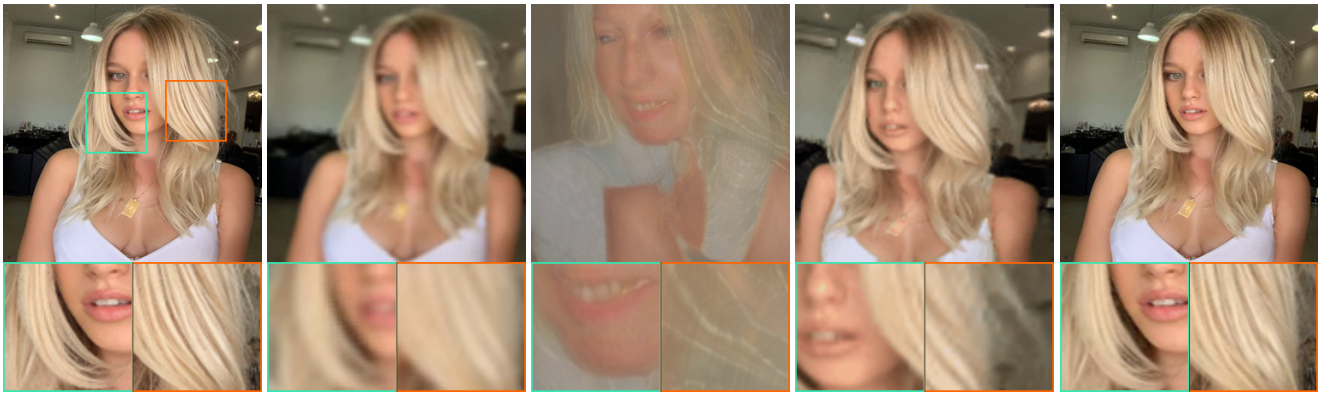


Figure 6: Qualitative Comparison results of our approach with recent video prediction method [32], GAN-based method [59], diffusion model [55]. In each image, we zoom in rectangle regions marked by red/green at the bottom.

Metric	FVD ↓	E_{warp} ↓
Chuang *et al*. [15]	1263.44	547.64
Halperin *et al*. [23]	1778.24	661.34
Endo *et al*. [19]	2026.59	1280.66
Qiu *et al*. [59]	1329.46	573.78
Hu *et al*. [32]	1301.96	969.96
Ni *et al*. [55]	1192.75	**319.05**
Ours	**1153.98**	521.96

Table 1: Quantitative comparison results on 114 portrait images of SketchHairSalon[80]. Our method outperforms state-of-the-art approaches.

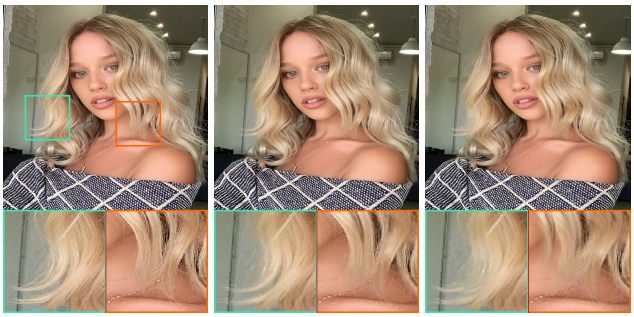


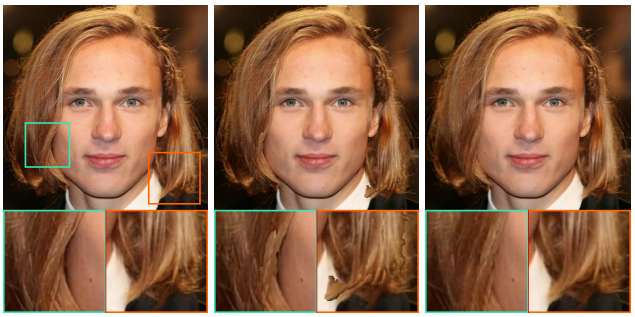
Figure 7: Effect of hair wisp extraction

extraction enables our method to generate various motions among wisps through extracting hair wisps, which is helpful to simulate complex and natural motions. Moreover, our hair wisp animation module effectively generates realistic motions for a hair wisp by building a physics-based model. As a result, our method generates videos with the highest quality compared with all baseline methods.

In addition, Fig. 6 and Tab. 1 compares our method with the state-of-the-art video prediction method [32], GAN-based method [59], and diffusion model [55]. They are the most recent single-image-to-video generation methods. As shown in Fig. 6 and Tab. 1, Qiu *et al*. [59] introduces noticeable distortions, while Hu *et al*. [32] creates unnatural motions, due to the challenges of hair animation. Similarly, Ni *et al*. [55] does not properly model complicated hair dynamics, leading to static scenes and introducing distortions in the left chin in Fig. 6. In contrast, our method achieves the best performance in Fig. 6 and Tab. 1, compared with these methods.

User Study. We conduct a subjective user study to evaluate our method. 18 subjects of various ages are invited to participate in our user study. Following [33, 70], we adopt paired comparison which is widely used to subjectively evaluate image/video quality of generated images/videos. For each subject, we display a still photo and two animated videos generated by different methods, where the photo is in the center and two videos are randomly presented side-by-side. All subjects have no prior knowledge of technical details. We asked each subject a question: *which video do you prefer by considering the original image?*

We compare our method with Halperin *et al*. [23], Endo *et al*. [19] and Chuang *et al*. [15] on 10 testing photos. In total, 81.5% of subjects are in favor of our method. These results indicate our method generates the most pleasing video, compared with all these state-of-the-art methods, although testing data contain complex backgrounds and various hairstyles and head poses.



Original w/o IHWE w FF Ours

Figure 8: Visual comparison of the contributions of our instance-based hair wisp extraction module

Method	FVD ↓	E_{warp} ↓
w/o IHWE w WE	1439.97	575.60
w/o IHWE w FF	1591.90	647.59
w/o HWA	1367.24	586.40
Ours	**1153.99**	**521.96**

Table 2: The ablation study results. Best in bold.

4.4. Ablation Studies

Instance-based Hair Wisp Extraction (IHWE.) We first validate the importance of hair wisp extraction in animating hair in a real image. We build a baseline named **w/o IHWE w WE** which removes our IHWE from our method and extracts the entire hair region instead. Fig. 7 shows **w/o IHWE w WE** creates unrealistic motions, which improperly enforces all hair wisps to undergo spatially consistent movements and fails to model the complexity of hair dynamics. Instead, thanks to our hair wisp extraction, our method well creates variations among hair wisps (see supplemental videos).

We then show the effectiveness of IHWE by the second of baseline named **w/o IHWE w FF**. **w/o IHWE w FF** employs a Flood Fill algorithm to extract hair wisps, instead of using IHWE. Tab. 2 shows **w/o IHWE w FF** generates videos with the lowest quality (FVD) and worst temporal consistency (E_{warp}). Fig. 8 and supplemental videos show **w/o IHWE w FF** tears many hair wisps, leading to noticeable discontinuity artifacts within a hair wisp.

Hair Wisp Animation (HWA). To evaluate the effectiveness of our hair wisp animation, we build a baseline named **w/o HWA** by treating a hair wisp as a strand and driving it by a mass-spring system, like synthetic-hair-based animation. **w/o HWA** achieves a worse FVD value than our method in Tab. 2, since such animation is unable to delicately model the inner dynamics of hair wisps (*e.g.* temporal shape deformation of a hair wisp). Fig. 9 also shows **w/o HWA** introduces flickering artifacts.

Original w/o HWA Ours

Figure 9: Visual comparison of the contributions of hair wisp animation module. w/o HWA results in distortions (region masked by cyan) and unrealistic motions (region masked by orange).

5. Conclusions

In this paper, we propose a novel approach that automatically animates hair in a still portrait photo without any user assistance. An instance-based hair wisp extraction module is proposed to extract hair wisps from an image, which facilitates the animation of hair and helps to generate complex hair motions. To train the instance segmentation model, we construct a hair wisp dataset containing real portrait photos and ground-truth annotations of hair wisps. Moreover, we introduce a hair wisp animation module that can create realistic motions for hair wisps based on physical models. Benefit from animated hair wisps, our method effectively converts diverse portrait photos containing various hairstyles and head poses into high-quality and high-resolution videos, but also enable the generated videos to provide an aesthetically-pleasing viewing experience without noticeable artifacts.

Acknowledgement

This work was supported by the KAUST Office of Sponsored Research through the Visual Computing Center (VCC) funding, as well as, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

[1] Cinemagraph. http://www.cinemagraph.com/. 1

[2] Ashampoo. Ashampoo cinemagraph. https://www.ashampoo.com/en-us/cinemagraph/. 2

[3] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (ToG)*, 36(6):1–13, 2017. 3

[4] Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. Automatic cinemagraph portraits. In *Com-*

puter Graphics Forum, volume 32, pages 17–25. Wiley Online Library, 2013. 1, 2

[5] Thabo Beeler, Bernd Bickel, Gioacchino Noris, Paul Beardsley, Steve Marschner, Robert W Sumner, and Markus Gross. Coupled 3d reconstruction of sparse facial hair and skin. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 2

[6] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14707–14717, 2021. 3

[7] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 5

[8] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7608–7617, 2019. 3

[9] Menglei Chai, Tianjia Shao, Hongzhi Wu, Yanlin Weng, and Kun Zhou. Autohair: Fully automatic hair modeling from a single image. *ACM Transactions on Graphics*, 35(4), 2016. 2, 4

[10] Menglei Chai, Lvdi Wang, Yanlin Weng, Xiaogang Jin, and Kun Zhou. Dynamic hair manipulation in images and videos. *ACM Transactions on Graphics (TOG)*, 32(4):1–8, 2013. 2, 6

[11] Menglei Chai, Lvdi Wang, Yanlin Weng, Yizhou Yu, Baining Guo, and Kun Zhou. Single-view hair modeling for portrait manipulation. *ACM Transactions on Graphics (TOG)*, 31(4):1–8, 2012. 2

[12] Johnny T Chang, Jingyi Jin, and Yizhou Yu. A practical model for hair mutual interactions. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 73–80, 2002. 4

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4, 6

[14] Byoungwon Choe and Hyeong-Seok Ko. A statistical wisp model and pseudophysical approaches for interactive hairstyle generation. *IEEE Transactions on Visualization and Computer Graphics*, 11(2):160–170, 2005. 2

[15] Yung-Yu Chuang, Dan B Goldman, Ke Colin Zheng, Brian Curless, David H Salesin, and Richard Szeliski. Animating pictures with stochastic motion textures. In *ACM SIGGRAPH 2005 Papers*, pages 853–860. 2005. 3, 6, 7, 8

[16] Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17622–17631, 2022. 7

[17] Gilles Daviet, Florence Bertails-Descoubes, and Laurence Boissieux. A hybrid iterative solver for robustly capturing coulomb friction in hair dynamics. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011. 2, 5

[18] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3742–3753, June 2021. 3, 7

[19] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: self-supervised learning of decoupled motion and appearance for single-image video synthesis. *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2019)*, 38(6):175:1–175:19, 2019. 1, 2, 3, 6, 7, 8

[20] Flixel. Cinemagraph pro. https://flixel.com/products/mac/cinemagraph-pro/. 2

[21] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (ToG)*, 37(6):1–12, 2018. 3

[22] Graphitii. Graphitii. https://graphitii.com/. 2

[23] Tavi Halperin, Hanit Hakim, Orestis Vantzos, Gershon Hochman, Netai Benaim, Lior Sassy, Michael Kupchik, Ofir Bibi, and Ohad Fried. Endless loops: Detecting and animating periodic patterns in still images. *ACM Trans. Graph.*, 40(4), Aug. 2021. 6, 7, 8

[24] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7854–7863, 2018. 3

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[26] Aleksander Holynski, Brian L. Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5819, June 2021. 1, 2, 3, 5

[27] Alexander Hornung, Ellen Dekkers, and Leif Kobbelt. Character animation from 2d pictures and 3d motion data. *ACM Transactions on Graphics (ToG)*, 26(1):1–es, 2007. 3

[28] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637*, 2021. 4

[29] 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

[30] 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

[31] Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. Single-view hair modeling using a hairstyle database. *ACM Transactions on Graphics (ToG)*, 34(4):1–9, 2015. 2

[32] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 3, 8

[33] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 8

[34] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. 2023. 4

[35] Wenzel Jakob, Jonathan T Moon, and Steve Marschner. Capturing hair assemblies fiber by fiber. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–9. 2009. 2

[36] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[37] Taewoo Kim, Chaeyeon Chung, Sunghyun Park, Gyojung Gu, Keonmin Nam, Wonzo Choe, Jaesung Lee, and Jaegul Choo. K-hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1299–1303. IEEE, 2021. 4

[38] Tae-Yong Kim and Ulrich Neumann. Interactive multiresolution hair modeling and editing. *ACM Transactions on Graphics (TOG)*, 21(3):620–629, 2002. 2

[39] Zhiyi Kuang, Yiyang Chen, Hongbo Fu, Kun Zhou, and Youyi Zheng. Deepmvshair: Deep hair modeling from sparse views. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 2

[40] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020. 7

[41] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7

[42] Mark Levoy. Area flooding algorithms. *Two-Dimensional Computer Animation, Course Notes 9 for SIGGRAPH*, 82, 1981. 4

[43] Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, and C-C Jay Kuo. Depth-aware stereo video retargeting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6517–6525, 2018. 6

[44] Bing Li, Chia-Wen Lin, Cheng Zheng, Shan Liu, Junsong Yuan, Bernard Ghanem, and C-C Jay Kuo. High quality disparity remapping with two-stage warping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2269–2278, 2021. 6

[45] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia*, 24:4077–4091, 2021. 3

[46] Peipei Li, Yinglu Liu, Hailin Shi, Xiang Wu, Yibo Hu, Ran He, and Zhenan Sun. Dual-structure disentangling variational generation for data-limited face parsing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 556–564, 2020. 4

[47] Kongming Liang, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Unifying visual attribute learning with object recognition in a multiplicative framework. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1747–1760, 2018. 3

[48] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[49] Zicheng Liao, Neel Joshi, and Hugues Hoppe. Automated video looping with progressive dynamism. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013. 2

[50] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5654–5663, 2019. 4

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

[52] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 5, 7

[53] Iacopo Masi, Joe Mathai, and Wael AbdAlmageed. Towards learning structure via consensus for face segmentation and parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5508–5518, 2020. 4

[54] 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

[55] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 18444–18455, 2023. 3, 8

[56] Tae-Hyun Oh, Kyungdon Joo, Neel Joshi, Baoyuan Wang, In So Kweon, and Sing Bing Kang. Personalized cinemagraphs using semantic understanding and collaborative learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5160–5169, 2017. 1, 2

[57] 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 Trans. Graph.*, 27(3):30, 2008. 2

[58] Eric Plante, Marie-Paule Cani, and Pierre Poulin. A layered wisp model for simulating interactions inside long hair. In *Computer Animation and Simulation 2001*, pages 139–148. Springer, 2001. 4

[59] Haonan Qiu, Yuming Jiang, Hang Zhou, Wayne Wu, and Ziwei Liu. Stylefacev: Face video generation via decomposing and recomposing pretrained stylegan3, 2022. 8

[60] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision, pages 12179–12188, 2021. 6

[61] J Reed and Elizabeth M Blunk. The influence of facial hair on impression formation. *Social Behavior and Personality: an international journal*, 18(1):169–175, 1990. 1

[62] Radu Alexandru Rosu, Shunsuke Saito, Ziyan Wang, Chenglei Wu, Sven Behnke, and Giljoo Nam. Neural strands: Learning hair geometry and appearance from multi-view images. *ECCV*, 2022. 2

[63] Andrew Selle, Michael Lentine, and Ronald Fedkiw. A mass spring model for hair simulation. In *ACM SIGGRAPH 2008 papers*, pages 1–11. 2008. 5

[64] Meng Sun, Allan D Jepson, and Eugene Fiume. Video input driven animation (vida). In *Computer Vision, IEEE International Conference on*, volume 2, pages 96–96. IEEE Computer Society, 2003. 3

[65] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2149–2159, January 2022. 6

[66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[67] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *European Conference on Computer Vision*, pages 258–274, 2020. 4

[68] James Tompkin, Fabrizio Pece, Kartic Subr, and Jan Kautz. Towards moment imagery: Automatic cinemagraphs. In *2011 Conference for Visual Media Production*, pages 87–93. IEEE, 2011. 1, 2

[69] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[70] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 8

[71] Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Chen Cao, Jason Saragih, Michael Zollhoefer, Jessica Hodgins, and Christoph Lassner. Neuwigs: A neural dynamic model for volumetric hair capture and animation. *arXiv preprint arXiv:2212.00613*, 2022. 5

[72] Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhöfer, Jessica Hodgins, and Christoph Lassner. Hvh: Learning a hybrid neural volumetric representation for dynamic hair performance capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6143–6154, June 2022. 2

[73] Kelly Ward, Florence Bertails, Tae-Yong Kim, Stephen R Marschner, Marie-Paule Cani, and Ming C Lin. A survey on hair modeling: Styling, simulation, and rendering. *IEEE transactions on visualization and computer graphics*, 13(2):213–234, 2007. 2

[74] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2019. 3

[75] WIKI. George raymond richard martin. https://en.wikipedia.org/wiki/GeorgeR.R.Martin. 1

[76] Sebastian Winberg, Gaspard Zoss, Prashanth Chandran, Paulo Gotardo, and Derek Bradley. Facial hair tracking for high fidelity performance capture. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 2

[77] Michael S Wogalter and Judith A Hosie. Effects of cranial and facial hair on perceptions of age and person. *The Journal of Social Psychology*, 131(4):589–591, 1991. 1

[78] Keyu Wu, Yifan Ye, Lingchen Yang, Hongbo Fu, Kun Zhou, and Youyi Zheng. 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*, pages 1526–1535, 2022. 2

[79] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 3

[80] Chufeng Xiao, Deng Yu, Xiaoguang Han, Youyi Zheng, and Hongbo Fu. Sketchhairsalon: Deep sketch-based hair image synthesis. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2021)*, 40(6):1–16, 2021. 4, 8

[81] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7632–7641, 2022. 3

[82] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[83] 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):1–11, 2014. 2

[84] Hang Yan, Yebin Liu, and Yasutaka Furukawa. Turning an urban scene video into a cinemagraph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 394–402, 2017. 2

[85] 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

[86] Mei-Chen Yeh and Po-Yi Li. A tool for automatic cinemagraphs. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1259–1260, 2012. 2

[87] Cem Yuksel, Scott Schaefer, and John Keyser. Hair meshes. *ACM Transactions on Graphics (TOG)*, 28(5):1–7, 2009. 2

[88] Meng Zhang, Pan Wu, Hongzhi Wu, Yanlin Weng, Youyi Zheng, and Kun Zhou. Modeling hair from an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 37(6):1–10, 2018. 4

[89] Meng Zhang and Youyi Zheng. Hair-gan: Recovering 3d hair structure from a single image using generative adversarial networks. *Visual Informatics*, 3(2):102–112, 2019. 2

[90] 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, 2012. 2

[91] Chen Zhao, Ali Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[92] Qingping Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4156–4165, 2022. 4

[93] Yi Zhou, Liwen Hu, Jun Xing, Weikai Chen, Han-Wei Kung, Xin Tong, and Hao Li. Hairnet: Single-view hair reconstruction using convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 2