HairStep: Transfer Synthetic to Real Using Strand and Depth Maps for Single-View 3D Hair Modeling

Yujian Zheng[1,2] Zirong Jin[2] Moran Li[3] Haibin Huang[3]
Chongyang Ma[3] Shuguang Cui[2,1] Xiaoguang Han[2,1*]
[1]FNii, CUHKSZ [2]SSE, CUHKSZ [3]Kuaishou Technology

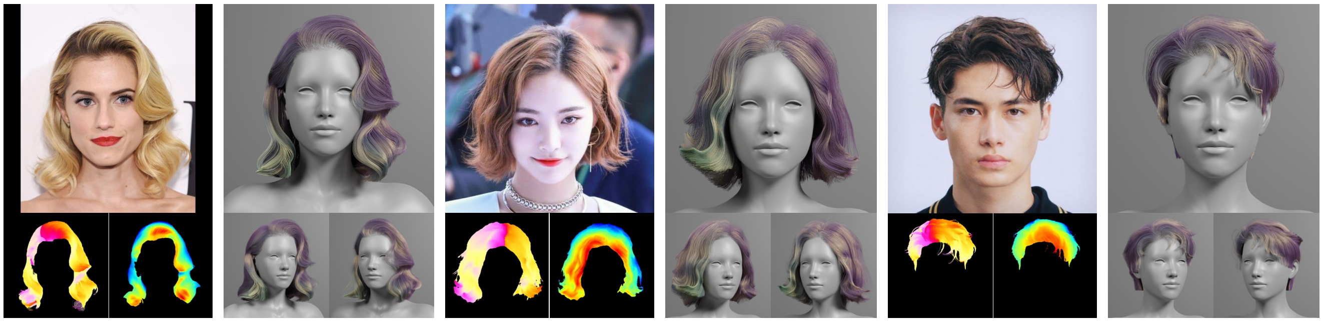


Figure 1. Given a single portrait image, we first convert it to an intermediate representation *HairStep* consisting of a strand map and a depth map (shown in the bottom left and right for each example), and then recover the corresponding 3D hair model at the strand level. Our *HairStep* is capable to bridge the domain gap between synthetic and real data and achieves high-fidelity hair modeling results.

Abstract

In this work, we tackle the challenging problem of learning-based single-view 3D hair modeling. Due to the great difficulty of collecting paired real image and 3D hair data, using synthetic data to provide prior knowledge for real domain becomes a leading solution. This unfortunately introduces the challenge of domain gap. Due to the inherent difficulty of realistic hair rendering, existing methods typically use orientation maps instead of hair images as input to bridge the gap. We firmly think an intermediate representation is essential, but we argue that orientation map using the dominant filtering-based methods is sensitive to uncertain noise and far from a competent representation. Thus, we first raise this issue up and propose a novel intermediate representation, termed as ***HairStep****, which consists of a strand map and a depth map. It is found that HairStep not only provides sufficient information for accurate 3D hair modeling, but also is feasible to be inferred from real images. Specifically, we collect a dataset of 1,250 portrait images with two types of annotations. A learning framework is further designed to transfer real images to the strand map and depth map. It is noted that, an extra bonus of our new dataset is the first quantitative metric for 3D hair modeling. Our experiments show that HairStep narrows the domain gap between synthetic and real and achieves state-of-the-art performance on single-view 3D hair reconstruction.*

1. Introduction

High-fidelity 3D hair modeling is a critical part in the creation of digital human. A hairstyle of a person typically consists of about 100,000 strands [1]. Due to the complexity, high-quality 3D hair model is expensive to obtain. Although high-end capture systems [9, 18] are relatively mature, it is still difficult to reconstruct satisfactory 3D hair with complex geometries.

Chai *et al*. [3,4] first present simple hair modeling methods from single-view images, which enable the acquisition of 3D hair more user-friendly. But these early systems require extra input such as user strokes. Moreover, they only work for visible parts of the hair and fail to recover invisible geometries faithfully. Recently, retrieval-based approaches [2, 10] reduce the dependency of user input and improve the quality of reconstructed 3D hair model. However, the accuracy and efficiency of these approaches are directly influenced by the size and diversity of the 3D hair database.

Inspired by the advances of learning-based shape recon-

[*] Corresponding author: hanxiaoguang@cuhk.edu.cn

struction, 3D strand models are generated by neural networks as explicit point sequences [45], volumetric orientation field [25, 29, 40], and implicit orientation field [36] from single-view input. With the above evolution of 3D hair representations, the quality of recovered shape has been improved significantly. As populating pairs of 3D hair and real images is challenging [45], existing learning-based methods [25, 29, 36, 39, 45] are just trained on synthetic data before applying on real portraits. However, the domain gap between rendered images (from synthetic hair models) and real images has a great and negative impact on the quality of reconstructed results. 3D hairstyles recovered by these approaches often mismatch the given images in some important details (e.g., orientation, curliness, and occlusion).

To narrow the domain gap between the synthetic data and real images, most existing methods [36, 37, 40, 45] take 2D orientation map [22] as an intermediate representation between the input image and 3D hair model. However, this undirected 2D orientation map is ambiguous in growing direction and loses 3D hints given in the image. More importantly, it relies on image filters, which leads to noisy orientation maps. In this work, we re-consider the current issues in single-view 3D hair modeling and believe that it is necessary to find a more appropriate intermediate representation to bridge the domain gap between real and synthetic data. This representation should provide enough information for 3D hair reconstruction. Also, it should be domain invariant and can be easily obtained from real image.

To address the above issues, we propose *HairStep*, a strand-aware and depth-enhanced hybrid representation for single-view 3D hair modeling. Motivated by how to generate clean orientation maps from real images, we annotate strand maps (i.e., directed 2D orientation maps) for real images via drawing well-aligned dense 2D vector curves along the hair. With this help, we can predict directed and clean 2D orientation maps from input single-view images directly. We also need an extra component of the intermediate representation to provide 3D information for hair reconstruction. Inspired by depth-in-the-wild [5], we annotate relative depth information for the hair region of real portraits. But depth learned from sparse and ordinal annotations has a non-negligible domain gap against the synthetic depth. To solve this, we propose a weakly-supervised domain adaptive solution based on the borrowed synthetic domain knowledge. Once we obtain the strand map and depth map, we combine them together to form *HairStep*. Then this hybrid representation will be fed into a network to learn 3D orientation field and 3D occupancy field of 3D hair models in implicit way. Finally, the 3D strand models can be synthesized from these two fields. The high-fidelity results are shown in Fig. 1. We name our dataset of hair images with strand annotation as *HiSa* and the one with depth annotation as *HiDa* for convenience.

Previous methods are mainly evaluated on real inputs through the comparison of the visual quality of reconstructed 3D hair and well-prepared user study. This subjective measurement may lead to unfair evaluation and biased conclusion. NeuralHDHair [36] projects the growth direction of reconstructed 3D strands, and compares with the 2D orientation map filtered from real image. This is a noteworthy progress, but the extracted orientation map is noisy and inaccurate. Moreover, only 2D growing direction is evaluated and 3D information is ignored. Based on our annotations, we propose novel and objective metrics for the evaluation of single-view 3D hair modeling on realistic images. We render the recovered 3D hair model to obtain strand and depth map, then compare them with our ground-truth annotations. Extensive experiments on our real dataset and the synthetic 3D hair dataset USC-HairSalon [10] demonstrate the superiority of our novel representation.

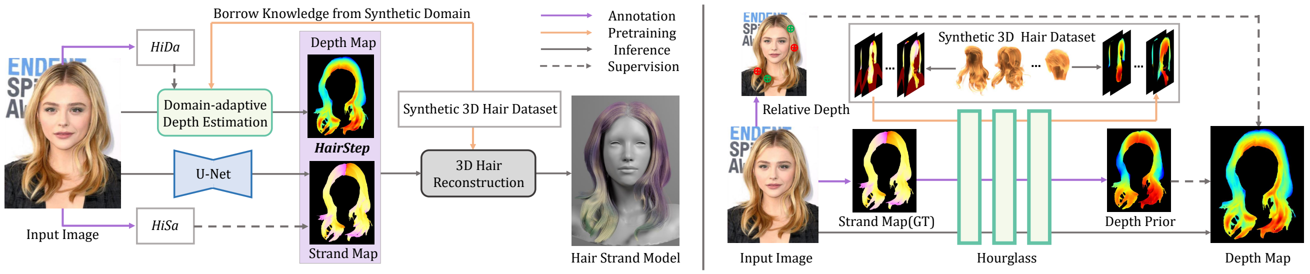
The main contributions of our work are as follows:

- We first re-think the issue of the significant domain gap between synthetic and real data in single-view 3D hair modeling, and propose a novel representation *HairStep*. Based on it, we provide a fully-automatic system for single-view hair strands reconstruction which achieves state-of-the-art performance.
- We contribute two datasets, namely *HiSa* and *HiDa*, to annotate strand maps and depth for 1,250 hairstyles of real portrait images. This opens a door for future research about hair understanding, reconstruction and editing.
- We carefully design a framework to generate *HairStep* from real images. More importantly, we propose a weakly-supervised domain adaptive solution for hair depth estimation.
- Based on our annotations, we introduce novel and fair metrics to evaluate the performance of single-view 3D hair modeling methods on real images.

2. Related Work

Single-view 3D hair modeling. It remains an open problem in computer vision and graphics to reconstruct 3D hair from a single-view input. Compared with multi-view hair modeling [18, 20, 39], single-view methods [4, 10, 36, 45] are more efficient and practical as multi-view approaches require carefully regulated environments and complex hardware setups. The pioneering single-view based methods [2–4, 10] typically generate a coarse hair model based on a database first, and then use geometric optimization to approximate the target hairstyles. The effectiveness of these approaches relies on the quality of priors and the performance is less satisfactory for challenging input.

Recently, with the rapid development of deep learning, several methods [25, 29, 36, 45] based on generative models have been proposed. HairNet [45] takes the orientation map as the input to narrow the domain gap between real



(a) The pipeline of single-view 3D hair reconstruction using *HairStep*

(b) Domain-adaptive depth estimation

Figure 2. Overview of our approach. (a) The pipeline of single-view 3D hair modeling with our novel representation *HairStep*. We collect two datasets *HiSa* and *HiDa*, and propose effective approaches for *HairStep* extraction from real images and finally realize high-fidelity 3D hair strand reconstruction. (b) Domain-adaptive depth estimation. We first pre-train the Hourglass on synthetic dataset, then generate depth priors as pseudo labels and finally obtain reasonable hair depth weakly-supervised by depth prior and annotated relative depth.

images and synthetic data, which enables the network to be trained with large-scale synthetic dataset. Hair-VAE [25] adopts a variational autoencoder to generate hair models from single-view input. Hair-GAN [40] introduces GAN based methods to the hair generation process. However, the hair models reconstructed by these methods tend to be coarse and over-smoothed, mainly due to the limited capacity of 3D neural network. To address this issue, Neural-HDHair [36] proposes a coarse-to-fine manner to obtain the high resolution 3D orientation fields and occupancy fields, enabling the GrowingNet to generate decent hair models.

Orientation maps for hair modeling. Due to the intrinsic elongated shapes of hair strands, it is intuitive to use 2D orientation maps and/or 3D orientation fields as intermediate representations to guide the modeling process. Existing image-based hair modeling methods typically apply Gabor filters of different directions to the input portrait and compute the local 2D orientation to follow the direction with the maximum filtering response [22, 23]. These 2D orientation maps are then converted into 3D orientation fields based on multi-view calibration information [9, 18, 19] or fed into neural network directly as auxiliary input for prediction of the 3D target hairstyle [36, 37, 40, 45]. However, 2D orientation maps based on image filtering operations suffer from input noise, which can be mitigated via additional smoothing or diffusion process at the expense of reduced accuracy [18, 19]. More importantly, these 2D orientation maps and 3D orientation fields do not distinguish between hair roots and tips from structure point of view. Addressing this kind of directional ambiguity requires additional input, such as user sketches [29] and physics based examples [9], which can be tedious or may not generalize well. Some methods [34] for 2D hair image generation are also based on orientation map.

Depth map estimation. Many data-driven methods [8, 11, 16, 28] using advanced techniques have achieved convincing performance on depth estimation. However, these approaches rely on dense depth labeling [6, 12, 13, 30], which is inaccessible for hair strands. Chen *et al*. [5] obviate the necessity of dense depth labeling by annotation of relative depth between sparse point pairs to help estimate depth map in the wild. However, there is no existing work to estimate depth map specifically for hair strands. Most 3D face or body reconstruction methods [27, 33, 35] only produce a coarse depth map of the hair region, which is far from enough for high-fidelity hair modeling.

3. *HairStep* Representation

The ideal way to recover 3D hair from single images via learning-based technique is to train a network which can map real images to the ground-truth 3D hair strands. But it is difficult and expensive to obtain ground-truth 3D hair geometries for real hair images [45]. [25] can only utilize a retrieval-based method [10] to create pseudo 3D hair models. Networks trained on such data can not produce 3D hairstyles aligned with given images, because it is hard to guarantee the alignment of retrieved hair with the input image. Due to the inherent difficulty of realistic hair rendering, existing methods [36, 37, 40, 45] take orientation maps instead of hair images as input to narrow the domain gap between real and synthetic data. However, orientation map obtained by image filters suffers from uncertain noise and is far from a competent intermediate representation. Hence, a better one is needed to bridge the significant gap.

We now formally introduce our novel representation *HairStep* for single-view 3D hair modeling. The overview of our method is shown in Fig. 2. We first give the definition of *HairStep* in Sec. 3.1, then describe how to obtain it from real images in Sec. 3.2 and Sec. 3.3. We describe how to use *HairStep* for single-view 3D hair modeling in Sec. 4.

3.1. Definition

Given a target image, we define the corresponding representation *HairStep* as $\mathbf{H} = \{\mathbf{O}, \mathbf{D}\}$, where $\mathbf{O}$ and $\mathbf{D}$ are the strand map and the depth map, respectively. The strand map $\mathbf{O}$ is formulated as an RGB image with a dimension of $W \times H \times 3$, where W and H are the width and the height of the target image. The color at a certain pixel $\mathbf{x}$ on the strand map is defined as

$$\mathbf{O}(\mathbf{x}) = (\mathbf{M}(\mathbf{x}), \mathbf{O}_{2\mathrm{D}}/2 + 0.5). \tag{1}$$

We use the red channel to indicate the hair mask with a binary map $\mathbf{M}$. We normalize the unit vector of projected 2D orientation $\mathbf{O}_{2\mathrm{D}}$ of hair growth at pixel $\mathbf{x}$ and represent this growing direction in green and blue channels. The depth map $\mathbf{D}$ can be easily defined as a $W \times H \times 1$ map where it represents the nearest distance of hair and the camera center in the camera coordinate at each pixel of hair region. Visual examples of *HairStep* are shown in Fig. 1.

Difference with existing representations. The existing 2D orientation map uses un-directed lines with two ambiguous directions [22] to describe the pixel-level hair growing in the degree of 180 while our strand map can represent the direction in the degree of 360 (see Fig. 3 (d-e)). NeuralHDHair [36] attempts to introduce an extra luminance map to supplement the lost local details in the real image. Unfortunately, there is a non-negligible domain gap between the luminance of synthetic and real images. Because it is highly related to the rendering scenarios such as lighting and material. Compared to the luminance map, our hair depth map only contains geometric information, which helps to narrow the domain gap of the synthetic and real images.

3.2. Extraction of Strand Map

To enable learning-based single-view 3D hair modeling, *HairStep* needs to be firstly extracted from both synthetic 3D hair data and real images for training and testing. For the synthetic data, we can easily obtain strand maps and depth maps from 3D strand models assisted by mature rendering techniques [17]. But it is infeasible to extract strand maps from real images via existing approaches. Thus, we use a learning-based approach and annotate a dataset *HiSa* to provide supervision.

***HiSa* dataset.** We collect 1,250 clear portrait images with various hairstyles from the Internet. The statistics of the hairstyles, gender and race are given in the supplementary material. We first hire artists to annotate dense 2D directional vector curves from the hair roots to the hair ends along the hair on the image (see the example in Fig. 3 (b)). On average, every hair image needs to cost about 1 hour of a skillful artist to draw about 300 vector curves. Once we obtain the dense strokes of the hair region, we convert them to a stroke map colored by the definition of Eq. (1), as shown in Fig. 3 (c). At last, we interpolate the colorized strokes within the mask of hair to obtain the ground-truth strand map (Fig. 3 (e)) of a given image. Thanks to the dense annotation, the holes are simple to be filled with ignorable loss of details. Compared with the undirectional orientation map extracted by Gabor filters (Fig. 3 (d)), our strand map is clean and can represent the growing direction without ambiguity.

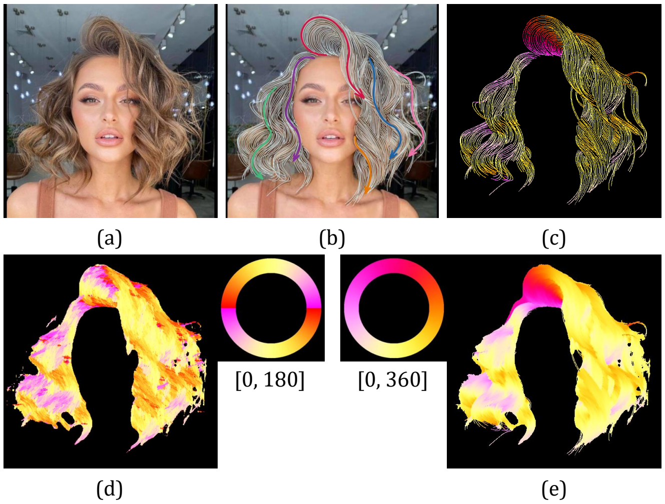


Figure 3. Obtaining strand map from vector strokes. (a) Portrait image. (b) Annotated vector strokes. (c) Colored strokes. (d) Orientation map extracted by Gabor filters. (e) Our strand map.

Strand map prediction. We consider the extraction of a strand map from a real image as an image-to-image translation task. We find that simply using an U-Net [24] can already achieve satisfactory results. Following standard settings, we use a pixel-wise L_1 loss and a perceptual loss against the ground-truth strand map $\mathbf{O}$, which is formulated as

$$\begin{aligned} L_{strand} = & \frac{1}{C \cdot \sum \mathbf{M}} \left\| \hat{\mathbf{O}} - \mathbf{O} \right\|_1 + \\ & \alpha \cdot \frac{1}{W_j H_j C_j} \left\| \phi_j(\hat{\mathbf{O}}) - \phi_j(\mathbf{O}) \right\|_2^2, \end{aligned} \tag{2}$$

where $\hat{\mathbf{O}}$ represents the predicted strand map and C represents the channel number of orientation map. The function $\phi_j(\cdot)$ represents the former j layers of pretrained VGG-19 [31] and we set j to be 35. W_j, H_j and C_j represent the shapes of output feature from $\phi_j(\cdot)$.

3.3. Domain-Adaptive Depth Estimation

It is not trivial to obtain the depth of hair from real images, because we cannot directly acquire the ground-truth depth annotation. Inspired by depth-in-the-wild [5], we annotate relative depth for the hair region of real images as weak labels. However, only constrained by the ordinal depth information, networks tend to generate unnatural depth maps. There is an obvious domain gap between learned depth from weak label and the synthetic depth used in the training, which leads to poor generalization when applying the trained model on real domain. Following the popular framework of domain adaptation based on pseudo labels [7, 14, 32, 41, 43, 44], we propose a domain-adaptive depth estimation method to reduce the gap of depth maps from real and synthetic data (see Fig. 2).

***HiDa* dataset.** We annotate depth relations for randomly selected pixel pairs in the hair region of each image among 1,250 portraits in *HiSa*. Different from depth-in-the-wild that only selects one pair per image, we annotate more than 140 pairs on average for each portrait which can give a more accurate and dense prediction. We first generate super-pixels within the hair region according to the ratio of the area of hair and face. We then randomly sample pixel pairs from all adjacent super-pixels and finally generate 177,074 pixel pairs in total for 1,250 real images. Two points in a pair are colored to red and blue, respectively. A QA program is designed to annotate the ordinal depth by showing one pair on the image each time and ask "which point in a pair of sampled pixels looks like closer to you, Red Point, Blue Point, or Hard to Tell?", following [5]. 12 well-trained workers are invited to annotate, which are split into three groups to ensure that every selected pair has been annotated three times by different groups. Finally 129,079 valid answers are collected (all groups give a certain relative depth, *i.e*. red or blue, and agree with each other). Our samplings takes a median of 4.6 seconds for a worker to decide, and three groups agree on the relative depth 72.9% of the time.

Learning depth map. We follow [5] to directly learn the mapping between the input image $\mathbf{I}$ and the output dense depth map $\mathbf{D}_r$ of the hair region through a Hourglass network [21], which is weakly supervised by our annotations. To train the network using ordinal labels of depth, we need a loss function that encourages the predicted depth map to agree with the ground-truth relations. We have found that the margin-ranking loss used in [15, 38, 42] works well in our task:

$$L_{rank} = \frac{1}{N}\sum_{i=1}^{N}\max(0, -(\mathbf{D}_r(p_1^i) - \mathbf{D}_r(p_2^i)) \cdot r^i + \varepsilon), \quad (3)$$

where p_1^i and p_2^i are pixels of the i_{th} annotated pair p^i, r^i is the ground-truth label which is set to 1 if p_1^i is closer otherwise -1. N represents the total number of sampled pairs in an image. ε is set to be 0.05, which gives a control to the difference of the depth values in p_1^i and p_2^i.

Domain adaptation. Although the ordinal label can provide local depth variation, it is a weak constraint which introduces ambiguity and leads to uneven solutions. The predicted depth map is usually unnatural and full of jagged artifacts (see the side views in Fig. 5). Applying this kind of depth to hair modeling often leads to coarse and noisy 3D shapes. To address above issues, we propose a weakly supervised domain-adaptive solution for hair depth estimation. We believe the knowledge borrowed from synthetic domain can help improve the quality of the learned depth.

Network trained with ordinal labels can not sense the absolute location, size and range of depth. The predicted depth has a major domain gap comparing to the synthetic depth map used in the training of 3D hair modeling. To give a constraint of the synthetic domain, we first train a network $Depth_{syn}$ to predict depth maps from strand maps on synthetic dataset by minimizing the L_1 distance between the prediction and the synthetic ground-truth. Then we input ground-truth strand maps of real images to $Depth_{syn}$ to query pseudo labels $\bar{\mathbf{D}}$ as depth priors. Note that directly applying this pseudo depth map to 3D hair modeling is not reasonable, because taking strand map as input can not provide adequate 3D information to the network. Jointly supervised by the depth prior and the weak-label of relative depth annotation, we predict decent depth maps which is not only natural-looking but preserves local relations of depth ranking. The loss function of the domain adaptive depth estimation is consisting of two parts, *i.e*., an L_1 loss against the pseudo label and the ranking loss defined in Eq. (3):

$$L_{depth} = \beta \cdot \left\|\mathbf{D}_r - \bar{\mathbf{D}}\right\|_1 + L_{rank}. \quad (4)$$

4. Single-View 3D Hair Modeling

Given the *HairStep* representation of a single-view portrait image, we further recover it to a strand-level 3D hair model. In this section, we first illustrate the 3D implicit hair representation, then describe the procedure of the reconstruction for hair strands.

4.1. 3D Hair Representation

Following NeuralHDHair [36], which is considered to be state-of-the-art in single-view hair modeling, we use implicit occupancy field and orientation field to represent 3D hair model in the canonical space of a standard scalp. The value of a point within the occupancy field is assigned to 1 if it is inside of the hair volume and is set to 0 otherwise. The attribute of a point in orientation field is defined as the unit 3D direction of the hair growth. The orientations of points outside of the hair volume are defined as zero vectors.

We use the same approach as [25] to extract the hair surface. During training, we sample large amount of points to form a discrete occupancy field. The sampling strategy follows [26] which samples around the mesh surface randomly and within the bounding box uniformly with a ratio of 1:1. For the orientation field, we calculate unit 3D orientations for dense points along more than 10k strands each model.

4.2. Strand Generation

To generate 3D stands from *HairStep*, we first train a neural network NeuralHDHair* following the method described by Wu *et al*. [36]. Taking our *HairStep* as input, the network can predict the implicit occupancy field and orientation field representing the target 3D hair model. Then we synthesis the hair strands adopting the growing method in [29] from the hair roots of the standard scalp.

The code of NeuralHDHair [36] has not been released yet and our own implementation NeuralHDHair* preserves the main pipeline and the full loss functions of NeuralHDHair, but has two main differences from the original NeuralHDHair. First, we do not use the sub-module of luminance map. The luminance has the potential to provide more hints for hair reconstruction, but suffers from the apparent domain gap between synthetic and real images, since it is highly related to the lighting. We attempt to apply the luminance map to the NeuralHDHair*, but it can only bring minor improvement. Second, we discard the GrowingNet of NeuralHDHair, since our work focuses on the quality of the reconstruction results instead of efficiency, while the GrowingNet is designed to accelerate the conversion from 3D implicit fields to hair strands. It maintains the same growth performance comparing to the traditional hair growth algorithm of [29], which is reported in [36].

5. Experiments

5.1. Datasets

We train the proposed method on USC-HairSalon [10] which is a publicly accessible 3D hairstyle database consisting of 343 synthetic hair models in various styles. We follow [45] to augment 3D hair data and select 3 random views each hair to generate corresponding strand maps and depth maps to form our *HairStep*. As for our real datasets *HiSa* and *HiDa*, we use 1,054 images with the resolution of 512×512 for training and 196 for testing. During training, we augment images and annotations by random rotating, scaling, translating and horizontally flipping.

5.2. Evaluation Metrics

Based on *HiSa* and *HiDa*, we propose two novel and fair metrics, i.e., *HairSale* and *HairRida*, to evaluate single-view 3D hair modeling results. We render the reconstructed 3D hair model to obtain strand map $\mathbf{O}_r$ and depth map $\mathbf{D}_r$, then compare them with our ground-truth annotations $\mathbf{O}_{gt}$ and $\mathbf{D}_{gt}$. Also, these two metrics can be applied to the evaluation on *HairStep* extraction.

HairSale. We first compute the mean angle error of growing direction called *HairSale* on rendered strand map, which ranges from 0 to 180. We define the *HairSale* as

$$\textit{HairSale} = \frac{1}{K}\sum_{\mathbf{x}^i \in U} \arccos(\mathcal{V}(\mathbf{O}_r(\mathbf{x}^i)) \cdot \mathcal{V}(\mathbf{O}_{gt}(\mathbf{x}^i))), \tag{5}$$

where U is the intersected region of rendered mask and the ground-truth. K is the total number of pixels in U. $\mathcal{V}(\mathbf{O}_r(\mathbf{x}^i))$ converts the color at pixel $\mathbf{x}^i$ of strand map $\mathbf{O}_r$ to an unit vector representing the growing direction.

HairRida. The *HairSale* only test the degree of matching in 2D. We also need a metric *HairRida* to measure the relative depth accuracy on *HiDa*, which is defined as

$$\textit{HairRida} = \frac{1}{Q}\sum_{i=1}^{Q} \max(0, r^i \cdot \text{sign}(\mathbf{D}_r(p_1^i) - \mathbf{D}_r(p_2^i))). \tag{6}$$

Note that we also calculate *HairRida* in the intersected region of rendered mask and the ground-truth. In addition, we provide the statistics of IoU for reference.

As for the evaluation of synthetic data, we follow [36] to compute the precision for occupancy field while using the L_2 error for orientation field.

5.3. Evaluation on HairStep Extraction

We first evaluate the effectiveness of our *HairStep* extraction method from real images. We found that simply applying a U-Net can already generate clean strand maps while Gabor filters suffer from uncertain noises (see Fig. 4). The *HairSale* computed on our predicted strand map is 12.3. As Gabor filters can only produce undirected orientation map, we convert the strand map to undirected map to calculate *HairSale* quantitatively for fair comparison. The *HairSale* on our results and Gabor's are 14.2 and 18.4 in undirected way, where our method performs 22.8% better. It's worth mentioning, the errors in undirected way are larger than in directed way, since the ambiguous bi-directed orientation leads to a worse measurement.

We evaluate the depth estimation using two metrics: the *HairRida* and a L_1 error against pseudo label (w/ or w/o normalization) to measure the difference between predicted depth and the synthetic prior. We compare the results of our domain-adaptive method $Depth_{DA}$ with the pseudo label $Depth_{pseudo}$ from synthetic domain, as well as the results of method only weakly supervised by ordinal label $Depth_{weak}$. The *HairRida* for $Depth_{pseudo}$, $Depth_{weak}$ and $Depth_{DA}$ are 80.47%, 85.17% and 85.20%, respectively. L_1 error against pseudo label (w/ ot w/o normalization) for $Depth_{weak}$ and $Depth_{DA}$ are 0.2470/3.125 and 0.1768/0.1188. Qualitative comparisons with different views of point cloud converted from depth maps are also

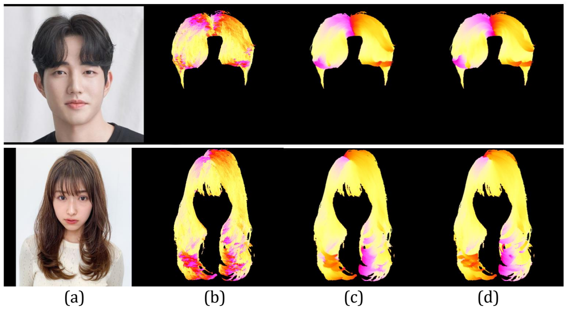


Figure 4. Qualitative comparisons on orientation/strand maps. (a) Input images; (b) undirected orientation maps from Gabor filters; (c) strand maps from our method; (d) ground-truth strand maps.

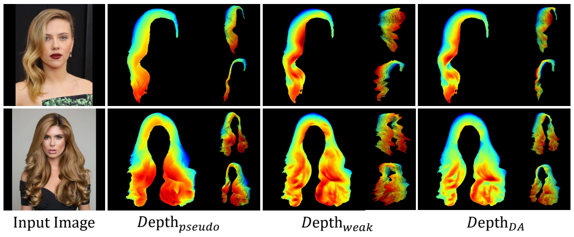


Figure 5. Qualitative comparisons on depth estimation.

shown in Fig. 5. The quantitative and qualitative give the same conclusion that our $Depth_{DA}$ is more competent to balance the local details of depth and the similarity of global shape to the synthetic prior. But the $Depth_{weak}$ is unnatural and full of serration-like artifacts. $Depth_{pseudo}$ suffers from the flat geometry, because the strand map can not provide strong 3D hints.

5.4. Comparisons

Comparisons on single-view hair modeling. We first compare the reconstruction results of NuralHDHair* (with the input of undirected orientation map), NuralHDHair* with our *HairStep*, HairNet [45], DynamicHair [37] and the original NeuralHDHair [36] in Fig. 6. We re-train HairNet and DynamicHair on our synthetic split, as they have not released pre-trained models. Based on a global latent code, HairNet and DynamicHair tend to generate coarse shapes while are not capable to reconstruct complex hairstyles. With the aid of the voxel-aligned feature and the implicit 3D representation, NeuralHDHair and NeuralHDHair* can produce decent results generally. However, it fails in the region with sharp variation of depth and the region with complicated pattern of hair growth (see Fig. 6). The reason could be that the undirected orientation map from Gabor filters can not provide clean and enough information for 3D hair modeling. Thanks to the novel representation *HairStep*, our results achieve the best.

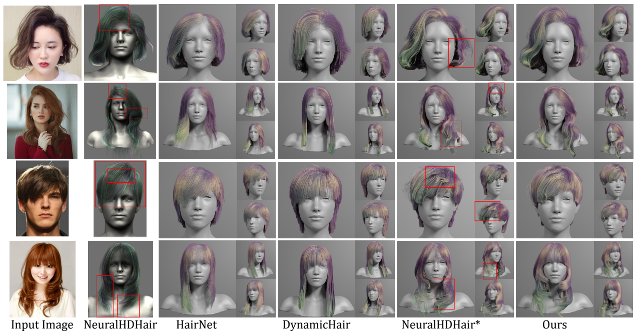


Figure 6. Comparisons with previous methods [36, 37, 45].

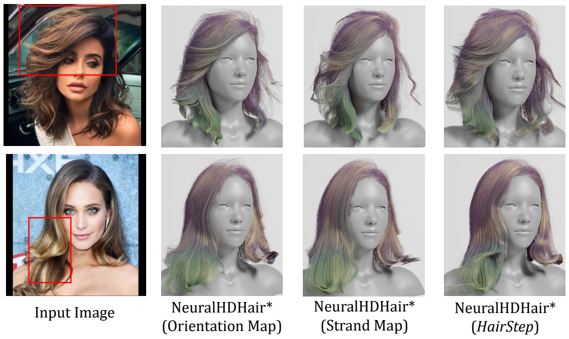


Figure 7. Qualitative evaluation results. From left to right: input images, results of NeuralHDHair*, results using our strand map based representation, and results of our full method, respectively.

Comparisons on representation. To evaluate the effectiveness of our *HairStep*, we compare it with strand map and existing orientation map [45] on three different frameworks, i.e. NeuralHDHair*, DynamicHair [37] and HairNet [45]. Quantitative comparisons on synthetic and real data are illustrated in Tab. 1 and Tab. 2, respectively. As shown in Tab. 1, our representation benefits all of these three methods on synthetic data. Since HairNet can only output explicit hair strands, we follow [45] to report mean square distance error in Tab. 1. In the evaluations on *HiSa* and *HiDa* of Tab. 2, it is proved that using our strand map achieves better alignment of hair growth than using previous orientation map [45] which suffers from ambiguous direction and image noise. The generalization ability of HairNet and DynamicHair is limited by the usage of global feature. Hence, directly concatenating depth information to the input does not seem helpful. Boosting by the full *HairStep*, there is an obvious improvement in depth accuracy on NeuralHDHair*. Qualitative comparisons shown in Fig. 7 yield the same conclusion, where *HairStep* performs the best in depth and preserves fine alignment of hair growth as same as strand map. Only applying orientation map leads to undesirable artifacts. Note that the depth accuracy of HairNet and DynamicHair in Tab. 2 is based on the low IoU,

Method	Orien. err. ↓	Occ. acc. ↑
NeuralHDHair* (Orientation map)	0.1324	82.59%
NeuralHDHair* (Strand map)	0.0722 (-41.7%)	84.18%
NeuralHDHair* (HairStep)	0.0658 (-50.3%)	86.77%
DynamicHair (Orientation map)	0.1352	78.19%
DynamicHair (Strand map)	0.1185 (-12.4%)	79.62%
DynamicHair (HairStep)	0.1174 (-13.2%)	79.78%
HairNet (Orientation map)	0.02349	/
HairNet (Strand map)	0.02206 (-6.1%)	/
HairNet (HairStep)	0.02184 (-7.0%)	/

Table 1. Quantitative comparisons on the USC-HairSalon dataset using different intermediate representations for NeuralHDHair*, DynamicHair [37] and HairNet [45].

Method	IoU ↑	*HairSale* ↓	*HairRida* ↑
NeuralHDHair* (Orientation map)	77.56%	19.6	70.67%
NeuralHDHair* (Strand map)	77.6%	**16** (-18.4%)	72.37%
NeuralHDHair* (HairStep)	77.22%	16.36 (-16.5%)	**76.79%**
DynamicHair (Orientation map)	56.39%	32.66	74.08%
DynamicHair (Strand map)	59.51%	26.53 (-18.8%)	73.42%
DynamicHair (HairStep)	59.14%	27.51 (-15.8%)	73.58%
HairNet (Orientation map)	57.15%	31.97	75.65%
HairNet (Strand map)	57.48%	28.6 (-10.5%)	74.81%
HairNet (HairStep)	57.01%	27.68 (-13.4%)	74.97%

Table 2. Quantitative comparisons on *HiSa* and *HiDa* of different intermediate representations for NeuralHDHair*, DynamicHair [37] and HairNet [45].

which is not comparable to NeuralHDHair*. In addition, we made a user study on 10 randomly selected examples involving 39 users for reconstructed results of NeuralHDHair* from three representations. 64.87% chose results from our *HairStep* as the best, while 21.28% and 13.85% for strand map and undirected orientation map.

5.5. Ablation Study

To better study the effect of each design in depth estimation on the final results, our representation is ablated with three configurations:

- $\boldsymbol{C_0}$: strand map + $Depth_{pseudo}$.
- $\boldsymbol{C_1}$: strand map + $Depth_{weak}$.
- $\boldsymbol{Full}$: strand map + $Depth_{DA}$.

Quantitative comparisons are reported in Tab. 3 and qualitative results are shown in Fig. 8. Our $\boldsymbol{Full}$ representation achieves the best result in depth accuracy and the decent alignment of hair growth. $\boldsymbol{C_0}$ suffers from the flat geometry of depth. Meanwhile, $\boldsymbol{C_1}$ can produce results with decent depth accuracy, but obtain a relatively larger difference on the alignment of hair growth than the $\boldsymbol{Full}$ representation.

6. Conclusion

In this work, we rethink the overall solution of single-view 3D hair modeling and argue that an appropriate intermediate representation for bridging the domain gap between synthetic and real data is essential. To this end, we propose a novel 3D hair representation *HairStep*, which consists of a strand map and a depth map, to narrow the existing domain gap. We also collect two datasets, i.e., *HiSa* and *HiDa*, with manually annotated strand maps and depth from real portrait images. These datasets not only allow the training of our learning based approach but also introduce fair and objective metrics to evaluate the performance of single-view 3D hair modeling. Extensive experiments on diverse examples demonstrate the effectiveness of our novel representation. Our method may fail on some rare and complex hairstyles, because the 3D network is basically overfitted on current synthetic datasets with limited amount and diversity.

Method	IoU ↑	*HairSale* ↓	*HairRida* ↑
$\boldsymbol{C_0}$	77.75%	16.03 (-18.2%)	73.57%
$\boldsymbol{C_1}$	77.11%	16.54 (-15.6%)	75.8%
$\boldsymbol{Full}$	77.22%	16.36 (-16.5%)	**76.79%**

Table 3. Quantitative ablation study about depth estimation.

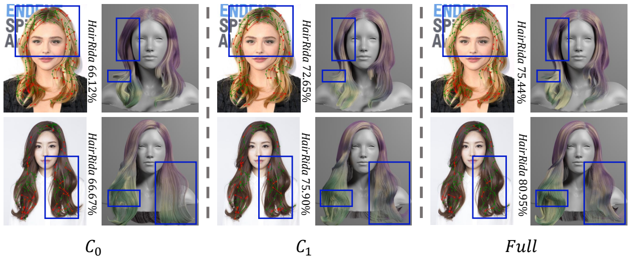


Figure 8. Qualitative ablation study results. Each pair from left to right: input image with the visualization of *HairRida*, where green/red line indicates right/wrong prediction of relative depth of two end point and the reconstructed 3D hair strand model (see Sec. 5.5 for detailed explanations).

Acknowledgement. The work was supported in part by NSFC with Grant No. 62293482, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). It was also partially supported by Shenzhen General Project with No. JCYJ20220530143604010.

References

[1] Yongtang Bao and Yue Qi. A survey of image-based techniques for hair modeling. *IEEE Access*, 6:18670–18684, 2018. 1

[2] 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. 1, 2

[3] 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. 1, 2

[4] 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. 1, 2

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 2, 3, 5

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 3

[7] Hehe Fan, Xiaojun Chang, Wanyue Zhang, Yi Cheng, Ying Sun, and Mohan Kankanhalli. Self-supervised global-local structure modeling for point cloud domain adaptation with reliable voted pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6377–6386, 2022. 5

[8] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*, pages 577–584, 2005. 3

[9] 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. 1, 3

[10] 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. 1, 2, 3, 6

[11] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2144–2158, 2014. 3

[12] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014. 3

[13] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 3

[14] Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2021. 5

[15] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 732–741, 2018. 5

[16] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015. 3

[17] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 4

[18] Linjie Luo, Hao Li, Sylvain Paris, Thibaut Weise, Mark Pauly, and Szymon Rusinkiewicz. Multi-view hair capture using orientation fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1490–1497. IEEE, 2012. 1, 2, 3

[19] Linjie Luo, Cha Zhang, Zhengyou Zhang, and Szymon Rusinkiewicz. Wide-baseline hair capture using strand-based refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 265–272, 2013. 3

[20] 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

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 5

[22] Sylvain Paris, Hector M Briceno, and François X Sillion. Capture of hair geometry from multiple images. *ACM transactions on graphics (TOG)*, 23(3):712–719, 2004. 2, 3, 4

[23] 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. 3

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[25] Shunsuke Saito, Liwen Hu, Chongyang Ma, Hikaru Ibayashi, Linjie Luo, and Hao Li. 3d hair synthesis using volumetric variational autoencoders. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018. 2, 3, 5

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

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

[28] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 3

[29] Yuefan Shen, Changgeng Zhang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Deepsketchhair: Deep sketch-based 3d hair modeling. *IEEE transactions on visualization and computer graphics*, 27(7):3250–3263, 2020. 2, 3, 6

[30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 3

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[32] Liangchen Song, Yonghao Xu, Lefei Zhang, Bo Du, Qian Zhang, and Xinggang Wang. Learning from synthetic images via active pseudo-labeling. *IEEE Transactions on Image Processing*, 29:6452–6465, 2020. 5

[33] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *European Conference on Computer Vision*, pages 246–264. Springer, 2020. 3

[34] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Lu Yuan, Sergey Tulyakov, and Nenghai Yu. Michigan: multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics (TOG)*, 39(4):95–1, 2020. 3

[35] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7750–7759, 2019. 3

[36] 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, 3, 4, 5, 6, 7

[37] 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, 3, 7, 8

[38] Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, Shuicheng Yan, M Shamim Hossain, and Ahmed Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9):1832–1842, 2016. 5

[39] 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. 2

[40] 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, 3

[41] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 5

[42] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 5

[43] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019. 5

[44] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 5

[45] 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, 3, 6, 7, 8