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

A. Dataset

Statistics of data distribution. To construct *HiSa* and *HiDa*, we collect 1,250 clear portrait images with various hairstyles from the Internet, where 80% are female and 20% are male. We classify the collected hairstyles into three classes, i.e., Short, Middle and Long, according to the position of their hair ends. If hair ends are above the mouth, the hairstyle will be classified as Short. If hair ends are below the shoulder, the hairstyle belongs to Long class. Otherwise, it is Middle. We collect 300 Short hair, 300 Middle hair and 650 Long hair. As for the curl type, the number of straight, wavy and curly are 210, 620, 420, respectively.

More examples about strand map annotation and depth pair sampling are shown in Fig. S1.

B. Implementation Details

We describe the details of our networks and training for *HairStep* extraction and 3D hair reconstruction in this section.

***HairStep* extraction.** We use the same U-Net in [2] to extract strand maps from real images with the resolution of 512×512. The network consists of an eight-layer encoder and an eight-layer decoder, where each layer downsamples/upsamples by a factor of 2 and skip connections are adopted between symmetric layers. We refer the readers to [2] for detailed designs. Training is conducted using a batch size of 16 for 50 epochs on 1 NVIDIA RTX3090Ti card for about 12 hours. The learning rate is 0.0003. During training, the loss weight α is set to 0.1.

We use the same Hourglass network in [1] to estimate depth maps for real images with the resolution of 512×512. The hourglass network is formed with four stacks, which consists of a series of convolutions, downsampling, upsampling and skip connections. Please refer to [1] for details. The network is trained with a batch size of 8 for 100 epochs on 2 NVIDIA RTX3090Ti cards for about 6 hours. The learning rate is 0.0003 and the loss weight β is set to 0.1.

3D Hair Reconstruction. We use the same structure as the IRHairNet in [3], where we first extract a $96 \times 128 \times 128 \times 64$ feature volume from the input representation resized to 256×256 via a U-Net combined with VIFu, then query coarse 3D occupancy field and orientation field with two MLPs. As for the fine module, we substitute the luminance map to the input representation resized to 1024×1024 and extract high-resolution occupancy field and orientation field via an hourglass network and two MLPs. Please refer to [3] for the details of network design. We follow [4] to combine the body mask to the mask channel of the strand map/orientation map rather than introducing a new channel. Note that our *HairStep* has one more depth channel than orientation map and strand map. Thus, the first layers of the encoders have 4 channels when using *HairStep*, while 3 channels when taking the strand map or orientation map as the input. Training is conducted using a batch size of 2 for 100 epochs on one NVIDIA RTX3090Ti cards for roughly 5-6 day. The learning rate is initialized set to be 0.0001, and decayed by a factor of 0.1 in the 60_{th} epoch.

C. Back views

Two examples of the back view are shown in Fig. S2 where the invisible parts tend to be smooth but still reasonable. This is because the 3D hair dataset provides shape priors.

D. Failure cases

As mentioned in the Conclusion, our method may fail on some rare and complex hairstyles, because the existing 3D hair datasets are with limited amount and diversity. For example, as shown in Fig. S4, our method does not work on hairstyles with braid (left) and complex curly pattern (right).

E. More Comparisons

Perceptual loss. We think the perceptual loss is necessary in strand map prediction. Although it cannot provide obvious quantitative improvement (w/ 14.2 *v.s.* w/o 14.1), it brings visually sharper local features (Fig. S3). Also, we made an extra experiment of 3D hair reconstruction on our method without perceptual loss. We found its *HairSale* and *HairRida* (16.51 and 75.3%) are worse than using perceptual loss (16.36 and 76.79%).

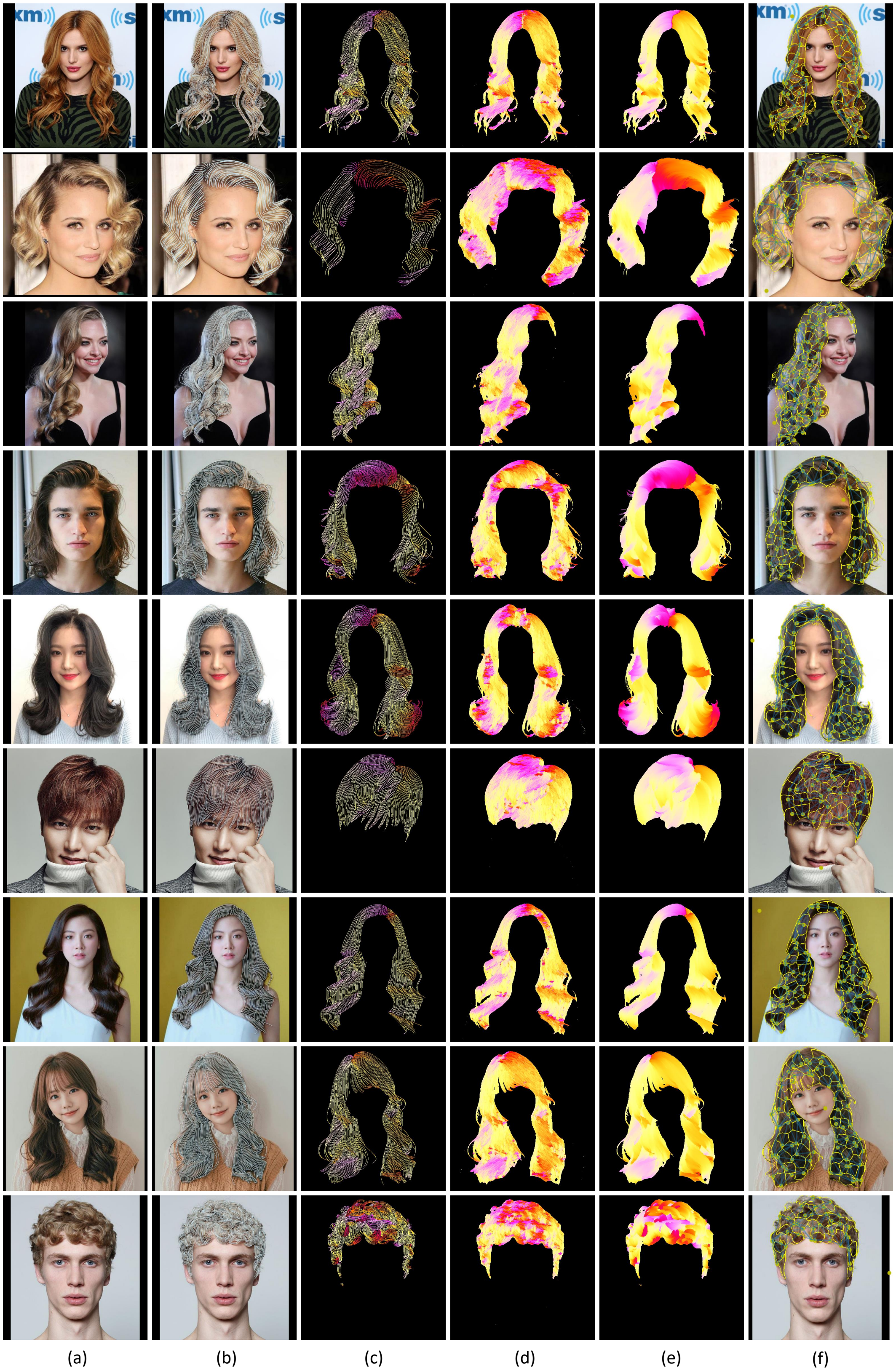

Figure S1. More examples for strand map annotation and depth pair sampling. From left to right: (a) collected images, (b) strokes drawn by artists, (c) colored strokes, (d) undirected orientation maps from Gabor filters, (e) strand maps and (f) super-pixels for depth pair sampling.

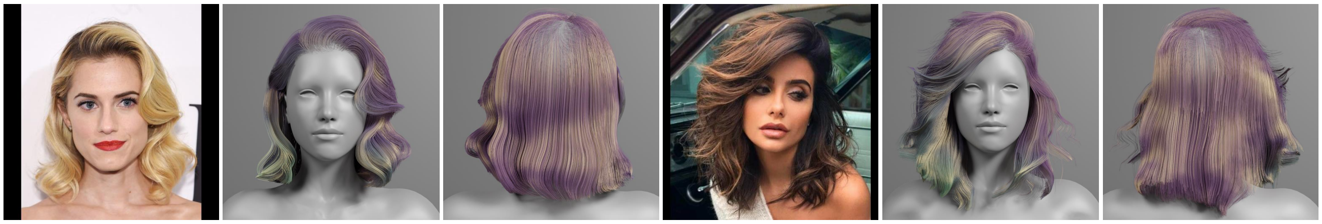

Figure S2. Examples with front and back view.

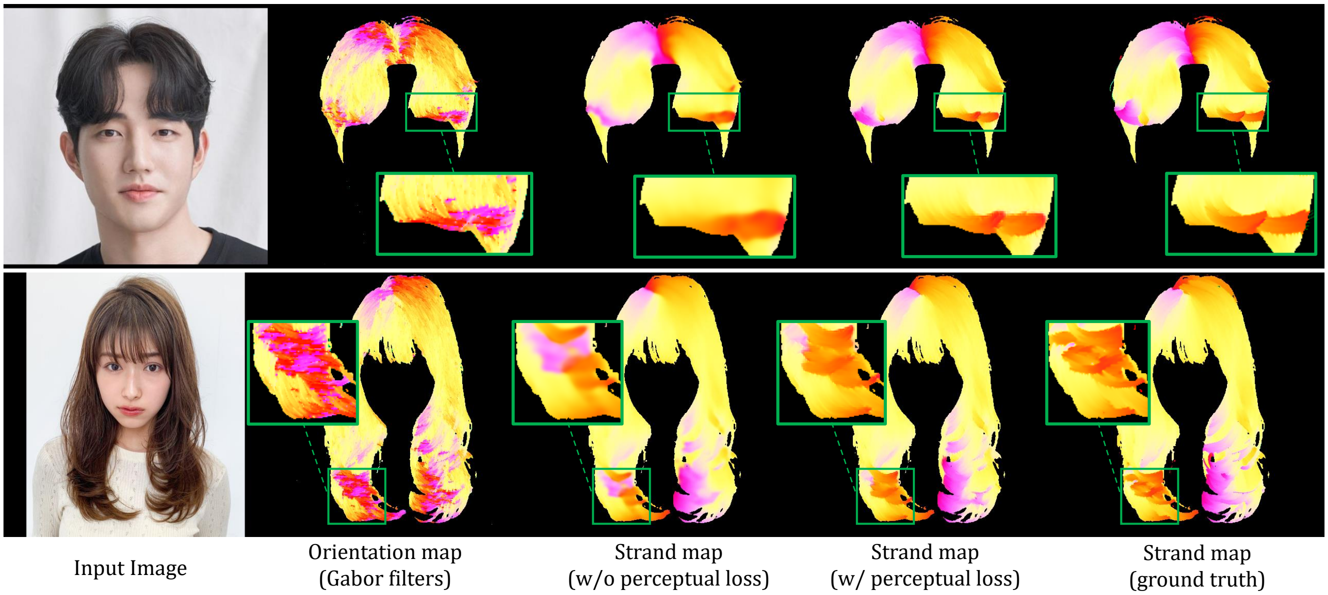
Input Image
Orientation map (Gabor filters)
Strand map (w/o perceptual loss)
Strand map (w/ perceptual loss)
Strand map (ground truth)

Figure S3. Qualitative comparisons on orientation/strand maps.

Qualitative comparisons for different representations. More qualitative comparisons for different representations are shown in Fig. S5 where using *HairStep* achieves the best results.

Qualitative comparisons for depth ablation. More qualitative comparisons for depth ablation are shown in Fig. S6 where our full model achieves the best accuracy in depth.

F. User Study

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. Fig. S7 and Fig. S8 provide the statistics of 3 different representations for each example.

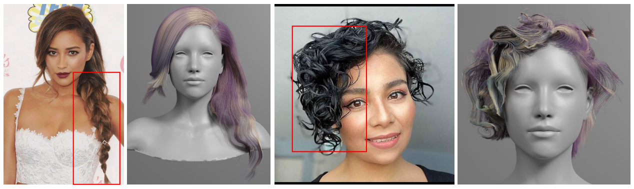

Figure S4. Failure cases.

References

[1] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 1

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[3] 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*

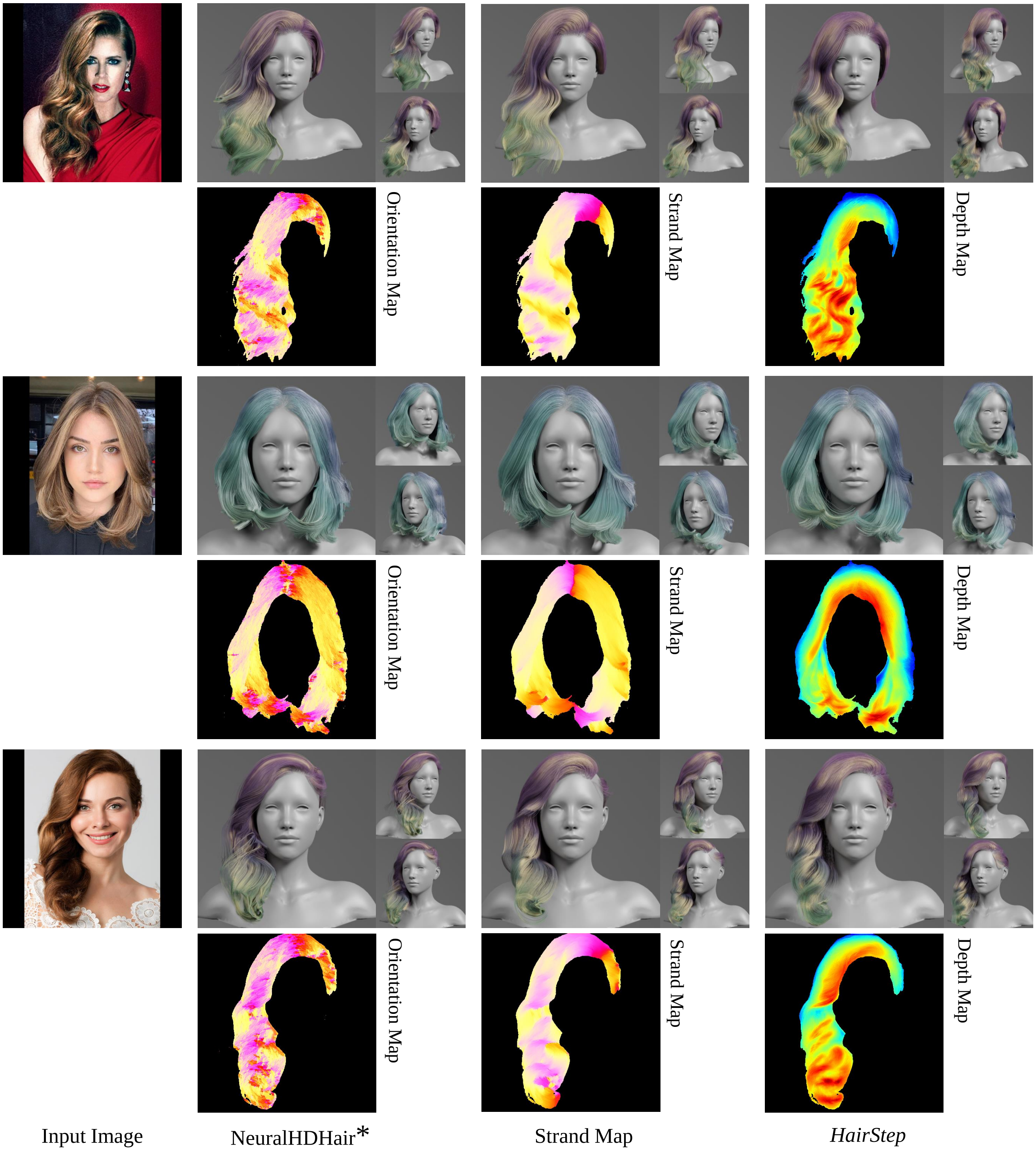

Figure S5. More qualitative comparisons for different representations. From left to right: input images, results of NeuralHDHair*, results using our strand map based representation, and results of our full method, respectively. Orientation maps from Gabor filters, predicted strand maps and depth maps are also shown under the reconstructed results.

Computer Vision and Pattern Recognition, pages 1526–1535, 2022. 1

[4] Yi Zhou, Liwen Hu, Jun Xing, Weikai Chen, Han-Wei Kung, Xin Tong, and Hao Li. Hairnet: Single-view hair reconstruc-

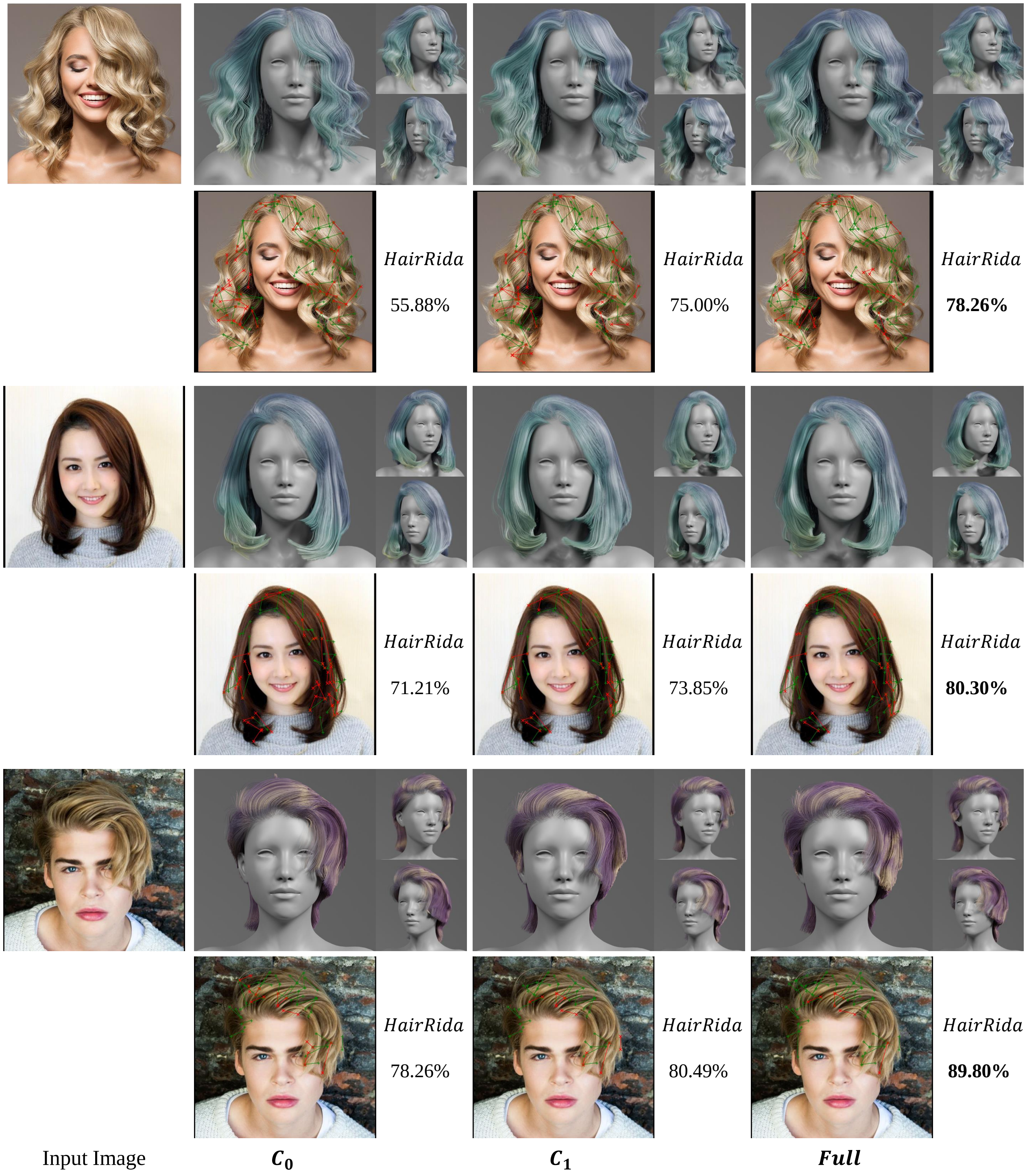

Figure S6. More qualitative comparisons for depth ablation. From left to right: input images, results of $\boldsymbol{C_0}$, $\boldsymbol{C_1}$ and our full method. We also visualize the *HairRida* below each reconstructed result, where green/red lines indicate right/wrong predictions.

tion using convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 1

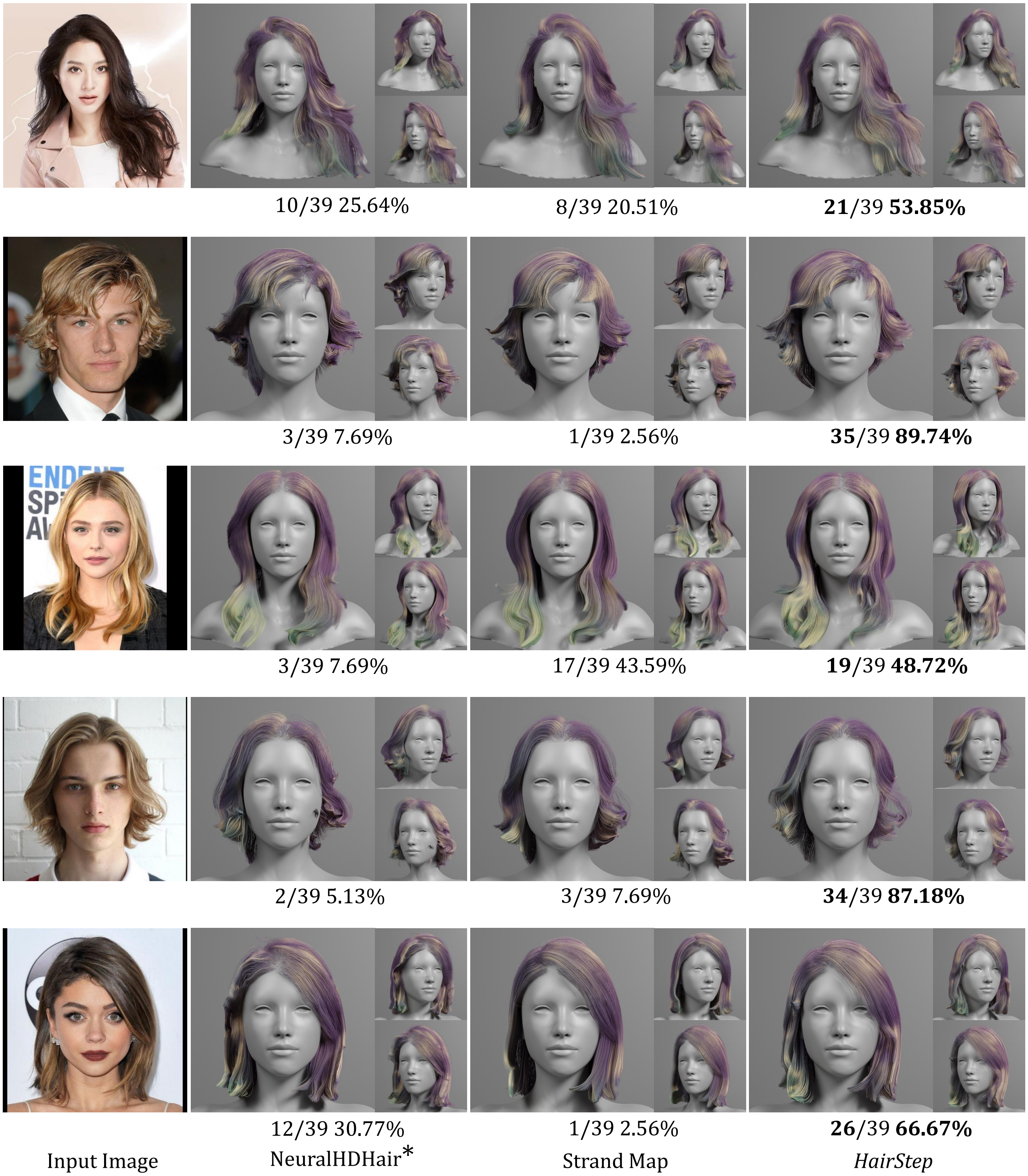

Figure S7. Examples for user study. From left to right: input images, results of NeuralHDHair*, results using our strand map based representation, and results of our full method, respectively. We also provide the statistics of 3 different representations for each example.

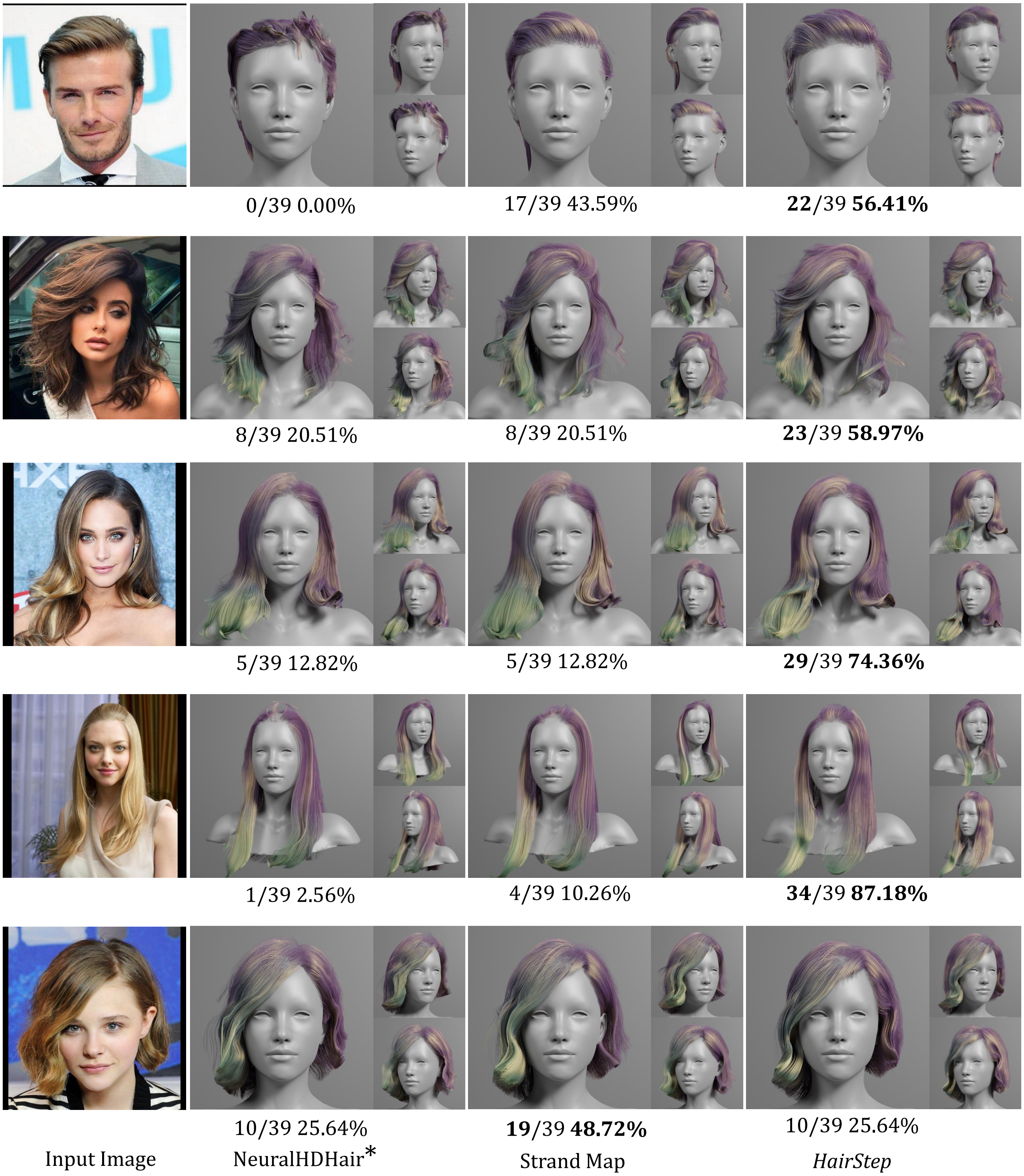

Figure S8. Examples for user study. From left to right: input images, results of NeuralHDHair*, results using our strand map based representation, and results of our full method, respectively. We also provide the statistics of 3 different representations for each example.