AnyBald: Toward Realistic Diffusion-Based Hair Removal In-The-Wild

Yongjun Choi[1*], Seungoh Han[1*], Soomin Kim[2], Sumin Son[2],
Mohsen Rohani[3], Edgar Maucourant[3], Dongbo Min[2], Kyungdon Joo[1†]

[1]Artificial Intelligence Graduate School, UNIST
[2]Ehwa Woman's University
[3]L'Oréal

[*]Equal contribution [†]Corresponding author

Abstract

We present `AnyBald`*, a novel framework for realistic hair removal from portrait images captured under diverse in-the-wild conditions. One of the key challenges is the lack of high-quality paired data, hindering real-world applicability. To address this, we construct a scalable data augmentation pipeline that synthesizes high-quality hair/non-hair image pairs capturing diverse real-world scenarios, enabling effective generalization with scalable supervision. Using this enriched dataset, we introduce a diffusion-based model with learnable text prompts that reformulates inpainting to work without explicit masks at inference. By doing so, ours can preserve semantics and produce natural results through implicit localization. Additionally, we introduce a regularization loss that guides the model to focus attention specifically on hair regions. Extensive experiments demonstrate that* `AnyBald` *outperforms in removing hairstyles while preserving identity and background semantics across various in-the-wild domains.*

1. Introduction

As digital human representation advances from coarse modeling to detailed components [18, 19], handling hair has become a key component [14], driving research in 3D reconstruction and hairstyle transfer [10, 22, 36] with applications in virtual try-on [2], digital avatars [28], and even in medical domain [11]. While prior works have mainly focused on hair reconstruction and transfer [3, 16, 23, 30, 34], natural hair removal remains under-explored, despite its importance for tasks like 3D face reconstruction and mesh recovery [26, 35], where hair's complex structure often causes occlusion and editing artifacts that degrade realism. A common solution is to remove the original hair before applying the target hairstyle [23, 34]. Here, we define hair removal as generating a clean bald version of a portrait while preserving all non-hair regions.

This problem remains challenging due to the inherent complex structure of hair, and collecting paired images is infeasible. Several studies have attempted to overcome the fundamental limitations of this task. HairMapper [24], for instance, learns a hair-removal path in the latent space of StyleGAN [9]. More recently, Stable-Hair [34] leverages a diffusion model to achieve more natural results. Yet, it still produces artifacts and often fails to preserve identity under challenging conditions such as non-frontal poses or unconventional compositions. To address these issues, we propose `AnyBald`, a unified framework for realistic hair removal in the wild.

`AnyBald` addresses data scarcity, hair complexity, and the need for mask-free inference by unifying data generation and model design. We first build a scalable augmentation pipeline that synthesizes diverse, high-quality paired data beyond prior low-quality, face-centered datasets [24], leveraging recent advances in controllable image generation [29, 31]. We then reformulate hair removal as an inpainting task, eliminating the need for region masks as input, along with learnable text embeddings in the BrushNet branch [8] and a text localization loss that guides attention only within hair regions. Together, the data generation and model training components form a unified pipeline that enables `AnyBald` to robustly remove hair from real-world portrait images while preserving semantic consistency in non-hair regions.

2. Method

We propose `AnyBald`, a unified framework for mask-free and realistic hair removal from in-the-wild portrait images. `AnyBald` tackles three key challenges in this task: the limited availability of diverse paired data, the structural complexity of hair, and the difficulty of obtaining accurate

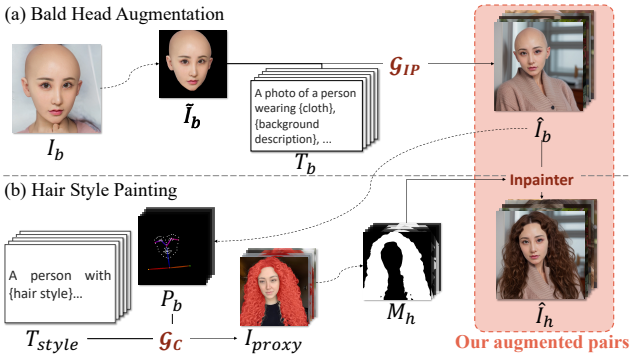


Figure 1. Overview of Paired Bald Augmentation. (a) From a bald image I_b, IP-Adapter is used to augment bald image $\hat{I}_b$ with processed reference bald $\tilde{I}_b$. (b) To synthesize the paired hair image $\hat{I}_h$, a proxy image I_{proxy} is first generated based on the head pose P_b of the augmented bald image $\hat{I}_b$, followed by the extraction of the aligned hair mask M_h using a face parser. The inpainting model utilizes M_h to generate the paired hair image $\hat{I}_h$ with photorealistic hairstyles.

masks. To address these issues, we integrate data augmentation, text-guided diffusion model, and spatial regularization.

`AnyBald` consists of three main components: (1) a generative augmentation pipeline that creates high-quality paired images of bald and haired versions under various poses and backgrounds (Sec. 2.1); (2) a dual-branch, mask-free diffusion model that uses learnable text prompts to selectively remove fine-grained hair while preserving identity and semantic content; and (3) a prompt-level attention loss that encourages the learnable prompt to attend more effectively to hair regions, enhancing spatial awareness of the target areas (Sec. 2.2). These components work together to enable robust and semantically consistent hair removal across a wide range of real-world conditions.

2.1. Paired Bald Augmentation

To overcome the limitations of existing bald image datasets, we propose a novel generative augmentation pipeline that constructs paired bald and haired images under diverse real-world conditions. Recent studies on hair-related editing tasks have utilized the non-hair-FFHQ dataset [24] to obtain bald images, which serve as a foundation for generating paired supervision [17, 34]. However, this dataset often contains residual hair, blurred boundaries, and low-quality artifacts, which can negatively affect the fidelity and consistency of the generated results. To address these limitations, we design a two-stage data augmentation process that first synthesizes clean bald images with varied poses, clothing, and backgrounds, and then aligns realistic hairstyles on them to form consistent training pairs (see Fig. 1), resulting pairs are shown in Fig. 2.

Bald Head Augmentation To enrich the diversity and realism, we perform a bald head augmentation that synthesizes cleaned bald images from noisy source bald data of non-hair-FFHQ. Given a source bald image I_b, we remove low-quality artifacts to obtain a cleaner reference image. Specifically, we use the face parsing model [27] to segment the face region and mask out residual hair and irrelevant areas (e.g., background and neck), producing a cleaned bald image $\tilde{I}_b$. To augment it $\tilde{I}_b$, we adopt a conditional generative model, IP-Adapter [29], which enables identity-preserving generation conditioned on both visual and textual inputs. This generator synthesizes a realistic bald image $\hat{I}_b$ by conditioning on both the cleaned bald $\tilde{I}_b$ and a text prompt T_b:

$$\hat{I}_b = \mathcal{G}_{IP}(\tilde{I}_b, T_b), \tag{1}$$

where $\mathcal{G}_{IP}$ represents the IP-Adapter. The prompt T_b represents key contextual variations, such as head pose (e.g., "side profile", "looking upward"), clothing style (e.g., "formal suit", "casual"), and background environment (e.g., "urban street", "indoor office"). This step enables us to produce a large-scale set of diverse and clean bald images with controlled variations in appearance and context (see Fig. 2(b)).

Figure 2. Samples of augmented pairs. (a) The source bald image I_b from non-hair-FFHQ. (b) $\hat{I}_b$ and (c) $\hat{I}_h$ indicate a generated pair of bald and haired images, respectively.

Hairstyle Painting From the augmented bald images, we construct bald/haired image pairs by aligning and inpainting realistic hairstyles on them. Precise alignment of the hairstyle on the bald head is essential to ensure consistency between bald and haired images. To this end, we present a hairstyle generation process that synthesizes an aligned hair mask using a pose-aware proxy image and then applies mask-guided hairstyle inpainting. To generate hair masks conditioned on the head pose, we adopt a pose-aware generative model based on ControlNet [31], which enables precise control through pose conditions extracted by Openpose [1]. First, we generate a proxy image I_{proxy} with a pose-conditioned generator $\mathcal{G}_C$:

$$I_{proxy} = \mathcal{G}_C(P_b, T_{style}). \tag{2}$$

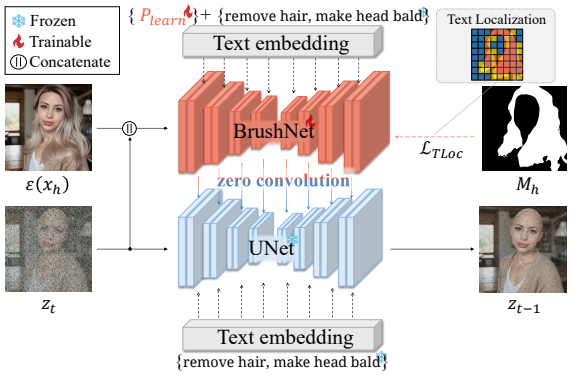


Figure 3. Overview of our training pipeline. The input image is encoded into a noisy latent z_t, which is concatenated with the $\mathcal{E}(x_h)$ and fed into the BrushNet branch $\mathcal{B}$. Features from each block are passed through zero convolution and injected into the corresponding U-Net block to predict the noise at step $t-1$. A learnable prompt P_{learn} is jointly optimized, while the hair mask M_h is used to constrain its spatial alignment with the latent features.

Figure 4. Qualitative results on CelebA and DeepFashion2.

where P_b denotes the pose representation given a augmented bald image $\hat{I}_b$, and T_{style} is a prompt that describes the desired hairstyle attributes. We utilize a fine-tuned version of ControlNet for pose conditioning, denoted as $\mathcal{G}_C$. Next, we extract a hair mask m from the generated proxy image I_{proxy} using a face parsing model [27], ensuring spatial alignment between the hairstyle and the target bald head. Finally, we inpaint the hairstyle using a mask-guided inpainting model [8], conditioned on the bald image $\hat{I}_b$ and the synthesized mask M_h, resulting in the final hair-augmented image $\hat{I}_h$.

Since obtaining bald images has been challenging in real-world, this high-quality generated set gives a novel alternative to the non-hair-FFHQ dataset. It not only alleviates the severe imbalance between male and female samples by increasing the proportion of female images, but also mitigates artifacts found in existing noisy bald images. We construct the `AnyBald` dataset, consisting of 11,404 images in total, split into 10,368 training images and 1,036 test images. Our augmentation pipeline and dataset will be publicly available for further exploration.

2.2. Hair Removal with Learnable Prompts

Model Architecture Inspired by prior mask-free approaches [20, 33], we develop a network that learns to identify and remove hair regions directly from paired data, avoiding coarse masks and preserving visual context despite hair's thin and complex structure. Our model adopts a dual-branch BrushNet pipeline (Fig. 3), where the upper branch $\mathcal{B}$ receives the concatenation of noisy and conditional latents $\mathcal{E}(x_h)$ from input image x_h through a VAE encoder $\mathcal{E}$. During training, $\mathcal{B}$ is updated while the U-Net remains frozen, and its cross-attended features with text condition c are injected into the U-Net via zero convolution. To enhance task awareness, we introduce a learnable prompt P_{learn} for hair removal, initialized from a context-aware prompt [37]. P_{learn} is prepended to the input prompt in $\mathcal{B}$, with its embedding τ_{learn} fine-tuned using the fixed text prompt *Remove hair, make head bald*, enabling it to specialize in hair removal.

Text Localization Loss Without explicit masks, P_{learn} may attend to irrelevant regions and cause artifacts. To address this, we use the hair mask M_h to define a text localization loss $\mathcal{L}_{TLoc}$ that concentrates attention on hair regions while suppressing it elsewhere [25, 33]. Notably, M_h is used only as indirect supervision during training and is not needed at inference. We define the text localization loss as:

$$\mathcal{L}_{TLoc} = \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(\alpha \, \mu_{i,k}^{\text{out}} - (1-\alpha) \, \mu_{i,k}^{\text{in}} \right) \quad (3)$$

where $\mu_{i,k}^{\text{out}} = \mu((1 - M_h^k) \odot A_i^{(k)})$ and $\mu_{i,k}^{\text{in}} = \mu(M_h^k \odot A_i^{(k)})$. Here, $A_i^{(k)}$ denotes the attention map of the i-th prompt token at layer k, n is the number of learnable prompt tokens, M_h^k is the interpolated hair mask, K is the number of cross-attention layers, and $\mu(\cdot)$ denotes the mean over attention heads and spatial tokens. We set α by the ratio of the hair-mask area to the image size to balance attention. The total loss is the standard latent diffusion loss $\mathcal{L}_{LDM}$ added with $\mathcal{L}_{TLoc}$.

3. Experiment

3.1. Experimental Setup

Baselines We compare with HairMapper (HM) and StableHair's Bald Converter (SH) using official checkpoints. For diffusion-based inpainting, we include ControlNet Inpainting (CNI) [31], BrushNet (Brush) [8], and PowerPaint v2

Method	SSIM ↑	LPIPS ↓	FID ↓	CLIP-I ↑
CNI-ft	0.801	0.193	38.30	0.918
Brush-ft	0.788	0.205	39.69	0.920
PPT-ft	0.771	0.193	30.47	0.953
HM	0.792	0.247	38.46	0.825
SH	0.787	0.205	29.94	0.925
Ours	**0.837**	**0.149**	**12.75**	**0.982**

Table 1. Quanitative comparison on the AnyBald test dataset. Bold and underlined metrics indicate the first and second best-performing methods, respectively.

Method	CelebA		DeepFashion2	
	AS ↑	IDS ↑	AS ↑	IDS ↑
HM	4.276	0.711	4.219	0.261
SH	4.675	0.602	4.512	0.286
Ours †(non-FFHQ)	4.747	0.624	4.562	0.235
Ours	**5.012**	**0.752**	**4.895**	**0.475**

Table 2. Quantitative comparison on two in-the-wild datasets. † indicates a trained version with non-hair-FFHQ.

	HM	SH	CNI-ft	Brush-ft	PPT-ft	Ours
Acc.	0.50	14.84	2.26	2.89	18.86	**60.62**
Pres.	0.50	10.18	4.27	3.52	22.89	**58.61**
Nat.	0.25	9.18	2.89	2.39	22.76	**62.51**

Table 3. User Study Results (%).

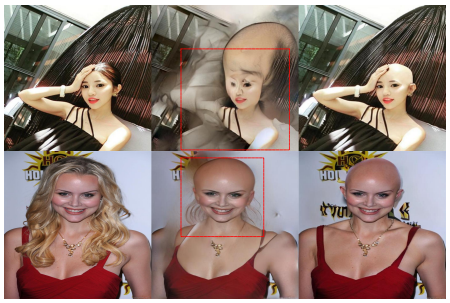
(PPT) [37], fine-tuned on our dataset for 10k steps with their default settings. All inpainting models use hair masks from a SegFormer-based face parser [5, 27].

Datasets For evaluation, we construct paired sets from our augmented dataset with GT bald images, since real-world pairs are rarely available. We also test generalization on in-the-wild datasets, using 1,867 images from CelebA (in-the-wild) [12] and 440 from DeepFashion2 [6] to evaluate hairstyle removal under unconstrained scenarios.

Evaluation Metrics SSIM [21] and LPIPS [32] evaluate structural and perceptual similarity, while FID [7] measures distributional distance between real and generated images. CLIP-I [13] captures semantic alignment via cosine similarity of CLIP embeddings. In the absence of GT bald images, we adopt Aesthetic Score (AS) [15] for visual quality and Identity Similarity (IDS) from ArcFace [4] embeddings.

3.2. Experimental Results

Qualitative evaluation Fig. 4 presents qualitative comparisons on the CelebA and DeepFashion2. HairMapper produces blurry backgrounds and bald artifacts in real images due to its strong dependence on the latent space of StyleGAN. While Stable-Hair works well on centered faces, it often struggles with complete hair removal in non-centered scenarios. Mask-based inpainting suffers from inconsistent shape and appearance within the masked area.



(a) Input (b) non-FFHQ (c) Ours

Figure 5. Ablation study. we compare the AnyBald dataset with the non-hair-FFHQ dataset.

Quantitative evaluation Table 1 shows the quantitative comparisons on the AnyBald test set compared to the all baselines including. The proposed method achieves the highest scores across various metrics, indicating superior generation quality. In Table 2, we evaluate the proposed method on two in-the-wild datasets, CelebA and DeepFashion2, comparing against existing hair removal models. To reduce the bias of our generated set, we also report the performance trained on the non-hair-FFHQ dataset, denoted as †. The proposed method achieves the highest AS and IDS across both datasets.

User Study We further evaluate quality through a user study with 53 participants on 30 samples (two sets of 15), drawn from unpaired datasets (CelebA, DeepFashion2, and collected web images). For each image, participants chose the best result in terms of (1)hair removal accuracy, (2) preservation of non-hair regions, and (3) visual naturalness. As shown in Table 3, AnyBald outperforms all baselines, consistently favored across diverse cases, while PowerPaint shows competitive results only with well-aligned masks and simpler hairstyles.

Ablation Study As shown in Fig. 5, we compare models trained on our dataset with non-hair-FFHQ. The model trained on non-hair-FFHQ produces blurry boundaries and poor identity preservation, while our dataset enables more realistic and generalizable hair removal under in-the-wild.

4. Conclusion

We introduce AnyBald, a method for realistic hair removal under various in-the-wild conditions. To support this task, we proposed a bald augmentation pipeline to construct a high-quality paired dataset called AnyBald dataset. Our pipeline integrates learnable prompts and text localization to remove hair regions effectively. Extensive experiments show that AnyBald outperforms prior methods both quantitatively and qualitatively, with strong human preference via user studies.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1): 172–186, 2019. 2

[2] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54 (4):1–41, 2021. 1

[3] Chaeyeon Chung, Sunghyun Park, Jeongho Kim, and Jaegul Choo. What to preserve and what to transfer: Faithful, identity-preserving diffusion-based hairstyle transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2582–2590, 2025. 1

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 4

[5] Jonathan Dinu. Face parsing model, 2022. 4

[6] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5337–5345, 2019. 4

[7] 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. 4

[8] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision (ECCV)*, pages 150–168. Springer, 2024. 1, 3

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 1

[10] Taewoo Kim, Chaeyeon Chung, Yoonseo Kim, Sunghyun Park, Kangyeol Kim, and Jaegul Choo. Style your hair: Latent optimization for pose-invariant hairstyle transfer via local-style-aware hair alignment. In *European Conference on Computer Vision (ECCV)*, pages 188–203. Springer, 2022. 1

[11] Wei Li, Alex Noel Joseph Raj, Tardi Tjahjadi, and Zhemin Zhuang. Digital hair removal by deep learning for skin lesion segmentation. *Pattern Recognition*, 117:107994, 2021. 1

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 4

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. 4

[14] Radu Alexandru Rosu, Keyu Wu, Yao Feng, Youyi Zheng, and Michael J Black. Difflocks: Generating 3d hair from a single image using diffusion models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10847–10857, 2025. 1

[15] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4

[16] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. Neural haircut: Prior-guided strand-based hair reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19762–19773, 2023. 1

[17] Kuiyuan Sun, Yuxuan Zhang, Jichao Zhang, Jiaming Liu, Wei Wang, Niculae Sebe, and Yao Zhao. Stable-hair v2: Real-world hair transfer via multiple-view diffusion model, 2025. 2

[18] Matthew A Turk, Alex Pentland, et al. Face recognition using eigenfaces. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991. 1

[19] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–I, 2001. 1

[20] Zhenchen Wan, Yanwu Xu, Dongting Hu, Weilun Cheng, Tianxi Chen, Zhaoqing Wang, Feng Liu, Tongliang Liu, and Mingming Gong. Mf-viton: High-fidelity mask-free virtual try-on with minimal input. *ArXiv*, abs/2503.08650, 2025. 3

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 4

[22] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18072–18081, 2022. 1

[23] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Gang Hua, and Nenghai Yu. Hairclipv2: Unifying hair editing via proxy feature blending. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 23589–23599, 2023. 1

[24] Yiqian Wu, Yong-Liang Yang, and Xiaogang Jin. Hairmapper: Removing hair from portraits using gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4227–4236, 2022. 1, 2

[25] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-

subject image generation with localized attention. *International Journal of Computer Vision (IJCV)*, 133(3):1175–1194, 2025. 3

[26] Liu Xiaoning, Wang Jie, Tuo Dongcheng, Jiang Jiaqi, Liang Zenglei, and Jiang Wenkai. A skull-face translation network used for generating faces from skulls. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1596–1599. IEEE, 2024. 1

[27] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2, 3, 4

[28] Jingrong Yang. A survey on hair modeling. *Highlights in Science, Engineering and Technology*, 115:512–526, 2024. 1

[29] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2

[30] Egor Zakharov, Vanessa Sklyarova, Michael Black, Giljoo Nam, Justus Thies, and Otmar Hilliges. Human hair reconstruction with strand-aligned 3d gaussians. In *European Conference on Computer Vision (ECCV)*, pages 409–425. Springer, 2024. 1

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 1, 2, 3

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 4

[33] Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and An-An Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 26399–26408, 2025. 3

[34] Yuxuan Zhang, Qing Zhang, Yiren Song, Jichao Zhang, Hao Tang, and Jiaming Liu. Stable-hair: Real-world hair transfer via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10348–10356, 2025. 1, 2

[35] Dapeng Zhao and Yue Qi. Generative landmarks guided eyeglasses removal 3d face reconstruction. In *International Conference on Multimedia Modeling*, pages 109–120. Springer, 2022. 1

[36] Yujian Zheng, Yuda Qiu, Leyang Jin, Chongyang Ma, Haibin Huang, Di Zhang, Pengfei Wan, and Xiaoguang Han. Towards unified 3d hair reconstruction from single-view portraits. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[37] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision (ECCV)*, pages 195–211. Springer, 2024. 3, 4