

Demographic Bias in Generative Data Augmentation: A Study of Makeup and Hairstyle Transfer Models

Waad A. Alotaibi
School of Computer Science and
Informatics
Cardiff University
Cardiff, United Kingdom
College of Computing and
Information Technology
Shaqra University
Shaqra, Saudi Arabia
alotaibiwa@cardiff.ac.uk

Sylwia Polberg-Riener
School of Computer Science and
Informatics
Cardiff University
Cardiff, United Kingdom
PolbergS@cardiff.ac.uk

Paul L. Rosin
School of Computer Science and
Informatics
Cardiff University
Cardiff, United Kingdom
RosinPL@cardiff.ac.uk

Abstract

Generative adversarial networks (GANs) are increasingly used for dataset augmentation through facial style transfer, yet they may replicate or amplify demographic bias present in their training data. We analyse bias in GAN-based makeup and hairstyle transfer models used for facial data augmentation. We evaluate three makeup models (BeautyGAN, PSGAN, EleGANt) and three hairstyle models (Barbershop, StyleYourHair, HairFastGAN) using a synthetic dataset of bald female faces balanced across four ethnicity-presenting groups (Asian, Black, Hispanic/Latino, White) and two expressions (neutral, smile). Qualitative and quantitative inspection reveals recurring failures including misalignment, photometric and chromatic artefacts, and unintended skin-tone drift, with the most severe degradations observed on Black-presenting faces and textured hairstyles. These findings highlight persistent generalisation gaps for underrepresented skin tones and hair textures, motivating evaluation protocols and training data that explicitly account for demographic equity.

CCS Concepts

• Computing methodologies → Computer vision.

Keywords

GANs, data augmentation, demographic bias, fairness in AI, facial style transfer, makeup transfer, hairstyle transfer, bias analysis, computer vision

ACM Reference Format:

Waad A. Alotaibi, Sylwia Polberg-Riener, and Paul L. Rosin. 2026. Demographic Bias in Generative Data Augmentation: A Study of Makeup and Hairstyle Transfer Models. In *The 8th International Conference on Control and Computer Vision (ICCCV 2026)*, April 03–05, 2026, Tsukuba, Japan. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3810417.3810425>



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICCCV 2026, Tsukuba, Japan*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2330-8/26/04
<https://doi.org/10.1145/3810417.3810425>

Introduction

Deep learning has become the state-of-the-art approach for many machine learning tasks, particularly in computer vision and natural language processing. However, training modern architectures typically requires large-scale data, and in many settings legal constraints, privacy concerns, and data scarcity make such collection impractical. Data augmentation is therefore widely used to increase variability and improve model generalisation. More recently, Generative Adversarial Networks (GANs) have been adopted for synthetic data generation [38]. Unlike conventional augmentation, GANs generate new samples that approximate the original data distribution and can be more effective in low-data regimes.

GAN-based augmentation has been applied across domains including medical imaging [11], anomaly detection [30], and emotion recognition [31]. Several studies report performance gains, particularly when training data are limited [15, 51], and GAN-augmented datasets are often described as robust and less sensitive to architecture and hyperparameters [17]. However, GAN augmentation can also replicate—and sometimes amplify—biases in the training distribution, thereby reproducing demographic disparities rather than correcting them. Jain et al. [22] showed that GAN-augmented engineering headshots were skewed toward masculine features and lighter skin tones based on human evaluation, while other work warns of biased associations in facial augmentation related to skin colour and racial representation [44]. Grissom et al. [16] further found that the StyleGAN3 discriminator assigns systematically higher scores to white faces, and that longer hair—particularly for Black individuals and men—tends to be penalised, suggesting structural preferences that can shape generation quality.

In this work, we investigate demographic bias in GAN-based style transfer models used for facial data augmentation. We evaluate three makeup transfer models and three hairstyle transfer models on a synthetic dataset of bald female faces balanced across four ethnicity-presenting groups and two expressions. By removing hair from the inputs, we enable controlled analysis of model behaviour while reducing confounds from occlusion. **Our main contributions are:**

- **Controlled bias-analysis protocol.** We curate a balanced bald-face evaluation set and apply fixed reference styles to enable matched comparisons across demographic groups.

- **Multi-metric evaluation.** We combine structured qualitative inspection with group-wise distribution-shift, photometric, and identity-preservation measures (FID, mid-face ΔL^* , and ArcFace cosine similarity) to assess augmentation-induced disparities.
- **Evidence of demographic performance gaps.** We identify consistent patterns of degradation for Black-presenting faces, including larger distribution shifts, unintended luminance changes, and frequent failures on textured and culturally specific hairstyles.

Related work

Image Data Augmentation

Data augmentation is widely used to improve model generalisation under limited data or class imbalance. Traditional approaches rely on geometric and photometric transformations, such as flipping, cropping, and colour perturbation, which were popularised by early deep vision models including AlexNet and VGGNet [28, 40]. More recently, generative augmentation using GANs has emerged as an alternative, synthesising new samples rather than transforming existing ones and proving particularly effective in low-data regimes [34, 49].

GAN-Based Augmentation for Facial Expression Recognition

Facial Expression Recognition (FER) frequently relies on data augmentation due to limited annotated datasets. In addition to conventional geometric augmentation, GAN-based methods have been shown to improve FER performance by generating synthetic expressions, with reported gains on benchmarks such as FER2013 and JAFFE [45, 51]. While effective for performance, these approaches rarely examine whether generative augmentation introduces or amplifies demographic bias—an issue central to our analysis.

Style Transfer Models for Data Augmentation

Style transfer has been widely adopted for data augmentation at both the image and feature levels. Early optimisation-based methods [12] were computationally expensive and often exhibited weak content–style alignment, motivating feed-forward transformation networks for efficient inference-time transfer [5, 20]. At the data level, style transfer diversifies the training distribution, while feature-level approaches perturb internal representations. Prior work improves robustness through style interpolation, feature mixing, and noise injection [19, 21].

Style transfer has also been explored for bias analysis and mitigation. Yucer et al. [46] transfer racial attributes while preserving identity, and [14] apply demographic style transfer to improve fairness in gender recognition. However, recent work shows that generative models trained on imbalanced data tend to reproduce or amplify demographic bias, with human evaluators often favouring outputs depicting white individuals [32].

Makeup transfer applies cosmetic styles from a reference image to a source while preserving facial identity. CycleGAN [49] enabled unpaired translation but lacked reference-specific control;

PairedCycleGAN [6] and BeautyGlow [7] addressed this by disentangling makeup and identity. BeautyGAN [29] improved visual fidelity through a dual-input/output architecture with a dedicated makeup loss but remained sensitive to spatial misalignment. PS-GAN [23] introduced attention-based pixel-wise correspondence to improve regional alignment, while EleGANt [43] employed a multi-scale pyramid encoder with self-attention for fine-grained, locally editable transfer. We therefore evaluate BeautyGAN, PS-GAN, and EleGANt as representative foundational, attention-based, and high-detail makeup transfer models.

Hairstyle transfer modifies hair colour, shape, and structure while preserving facial identity and background, with applications in virtual try-on and digital content creation. Advances in GAN architectures [25] have enabled increasingly realistic hair synthesis [24, 35]. Existing methods are commonly categorised as optimisation-based or encoder-based. Optimisation-based approaches refine latent codes using task-specific losses: Barbershop [50] controls structure and style in StyleGAN’s FS space but struggles with large pose variations, while StyleYourHair [27] improves robustness via local style matching and pose alignment; LOHO [37] and HairManip [48] further increase control through hair decomposition and disentanglement. Encoder-based methods prioritise inference efficiency; HairFastGAN [35] employs task-specific encoders for alignment, shape, colour, and refinement, enabling fast high-resolution synthesis, though challenges remain for diverse poses and complex textures. Accordingly, we analyse Barbershop, StyleYourHair, and HairFastGAN as representative optimisation-based and encoder-based hairstyle transfer models.

Dataset and Methodology

We construct a balanced evaluation set of 320 synthetic female images selected from 10,000 AI-generated photos released by Generated Photos [13]. All images are model-released and royalty-free. The final set is balanced across four *ethnicity-presenting* appearance categories (Asian, Black, Hispanic/Latino, White) and two facial expressions (neutral, smiling), with 40 identities per category per expression. We focus on female faces because current makeup and hairstyle transfer benchmarks and pretrained models are largely developed and evaluated on female data.

Appearance-based labels. Ethnicity-presenting categories are appearance-based labels provided by the source platform and are not self-identified or ground-truth ethnicity. To reduce ambiguous cases, we apply a label-consistency filtering step using DeepFace [39], excluding images where the provider label did not match the majority prediction of three DeepFace race classifiers. We then sample exactly 80 images per category. These labels enable controlled group-wise analysis but should not be interpreted as demographic ground truth. **Preprocessing (hair removal).** To standardise inputs and minimise hairstyle occlusions, all images are processed uniformly with HairMapper [42] to remove existing hair. The resulting bald images provide controlled inputs for evaluating demographic disparities introduced by style transfer models. Although preprocessing can introduce subtle bias, careful method selection and manual inspection revealed no systematic artefacts.

Models and evaluation protocol. We evaluate three makeup transfer models (BeautyGAN, PSGAN, EleGANT) on the full bald-face dataset (4 ethnicity-presenting groups $\times$ 2 expressions $\times$ 40 images = 320 sources). For hairstyle transfer, we assess Barbershop, StyleYourHair, and HairFastGAN on the Black-presenting subset (40 neutral + 40 smiling = 80 sources) using five reference hairstyles. This yields 2,560 makeup outputs per model (320×8 styles) and 400 hairstyle outputs per model (80×5 styles). Outputs are assessed via structured qualitative inspection using a fixed checklist covering alignment, color/texture fidelity, preservation of facial attributes, artefacts, and overall realism. Due to space constraints, figures show a stratified subset of 32 sources; full grids are provided in the supplementary material.

Distribution shift (FID). We report Fréchet Inception Distance (FID) [18] to quantify distribution shift introduced by makeup transfer. For each model and ethnicity-presenting group, we compare bald sources ($N=80$) to generated outputs pooled across eight reference styles ($N=640$), using Inception-v3 pool3 features after resizing images to 256×256 . We interpret FID as a relative cross-group shift metric rather than a perceptual quality score. **Photometric preservation (ΔL^*).** To quantify unintended global brightening or darkening, we measure mid-face luminance shift in CIELAB space. Using RetinaFace [39], we compute mean $L^* \in [0, 100]$ over a mid-face crop defined within the detected face bounding box and define $\Delta L^* = \bar{L}^*_{\text{gen}} - \bar{L}^*_{\text{src}}$, where positive values indicate facial brightening. We summarise both signed ΔL^* (direction) and $|\Delta L^*|$ (magnitude) per group and model. We include ΔL^* because unintended global brightening or darkening is a common failure mode in makeup transfer, and mid-face luminance provides a simple, model-agnostic measure of skin-tone perturbation that complements FID. **Identity preservation (ArcFace[10]).** We additionally quantify identity preservation under makeup transfer using cosine similarity between ArcFace face embeddings extracted from the bald source and the corresponding makeup output (higher is better). We report group-wise median and mean similarity over all matched outputs ($N=640$ per group). **Perceptual similarity (LPIPS).** We measure perceptual similarity to the reference hairstyle using LPIPS [47]. For each reference and expression, we detect the face and extract a hair-region ROI defined by fixed proportional offsets from the detected face bounding box, covering the hair and hairline. LPIPS is computed between the cropped reference and corresponding generated output (lower indicates greater perceptual similarity). Results are interpreted comparatively across models under fixed references.



Figure 1: Sample bald female faces from the synthetic dataset, balanced across four ethnicity-presenting groups and two expressions. These images serve as standardised inputs for style transfer evaluations.

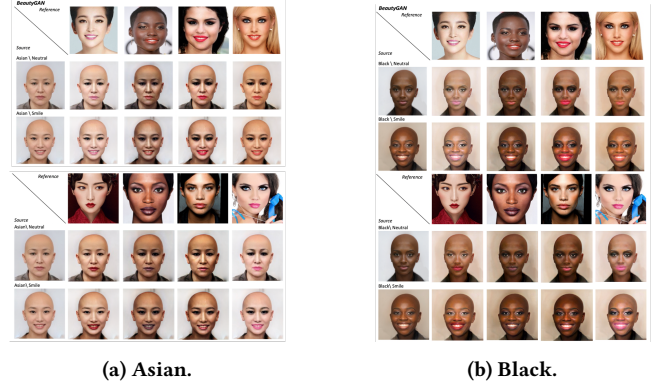


Figure 2: Makeup transfer results using BeautyGAN on Asian and Black.

Results

Makeup Transfer Results

We evaluated three makeup transfer models—BeautyGAN [29], PSGAN [23], and EleGANT [43]—on the synthetic bald female dataset to assess their suitability for demographically inclusive data augmentation. Reference images were drawn from the MT-Dataset [29] and public web sources, covering diverse skin tones and makeup intensities. Representative results across four ethnicity-presenting groups (Asian, Black, Hispanic/Latino, White) and two expressions (neutral, smiling) are shown in Figures 2–7.

BeautyGAN. Overall, BeautyGAN produces visually coherent makeup transfer when the source and reference share similar skin tones, but robustness decreases under cross-skin-tone transfer and expression variation. **Lips.** Lipstick is typically well aligned and saturated; however, smiling expressions often introduce edge blur and occasional colour spill onto teeth, especially for darker skin tones. **Eyes.** Eyeshadow placement is usually correct, though dramatic reference styles can appear smudged with reduced definition. **Skin tone (foundation).** Foundation transfer is smooth for same-tone references, while cross-tone references can introduce brightening, dull overlays, or desaturation—most pronounced for Black faces. **Expression effect.** Smiling amplifies lip misalignment and edge artefacts across all groups.

PSGAN. Overall, PSGAN exhibits unstable and highly group-dependent behaviour, with frequent failures in colour mapping and foundation adaptation. **Lips.** Lip transfer is inconsistent, with partial omission, bleeding, or over/under-saturation that becomes more noticeable in smiling expressions. **Eyes.** Eye makeup structure is often preserved, but saturation is uneven and hues frequently drift from the reference. **Skin tone (foundation).** Foundation is the weakest component, producing ashy overlays, patchiness, and over-darkening—particularly when source and reference skin tones differ. **Expression effect.** Smiling exacerbates lip bleeding and misalignment across all groups.

EleGANT. Overall, EleGANT delivers high-detail transfer on lighter skin tones but degrades sharply on darker complexions, revealing uneven demographic generalisation. **Lips.** Lip colour and shape transfer is accurate on Asian- and White-presenting

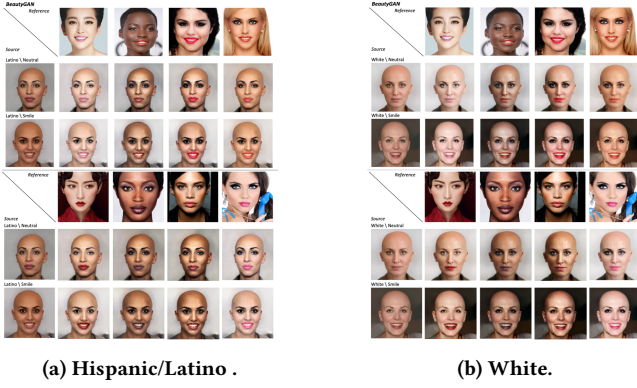


Figure 3: Makeup transfer results using BeautyGAN on Hispanic/Latino and White.

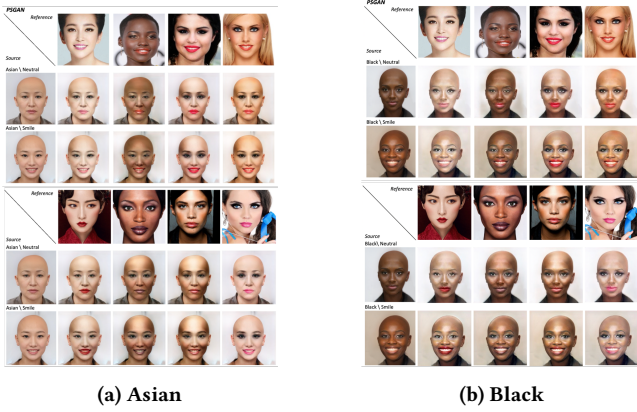


Figure 4: Makeup transfer results using PSGAN on Asian and Black.

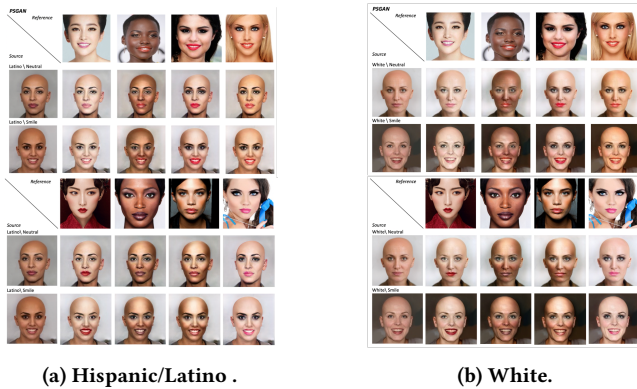


Figure 5: Makeup transfer results using PSGAN on Hispanic/Latino and White.

faces, but frequently bleeds or misaligns on Black-presenting faces. **Eyes.** Eye makeup appears crisp on lighter tones, while becoming smudged, pixelated, or oversaturated on darker complexions. **Skin**

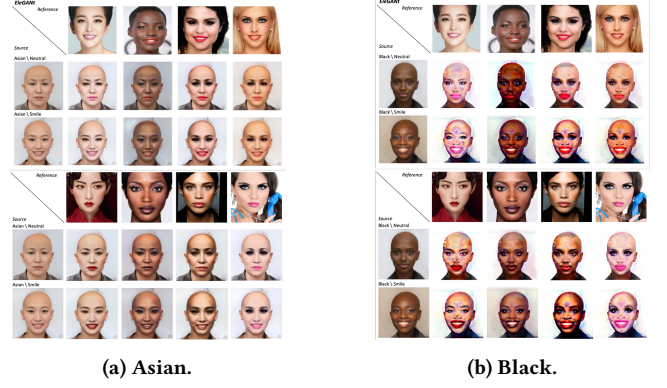


Figure 6: Makeup transfer results using EleGANT on Asian and Black.

tone (foundation). Foundation often appears as a global tonal overlay rather than subtle contouring, leading to unnatural darkening and chromatic noise—most evident on Black-presenting faces. **Expression effect.** Expression changes are secondary to skin-tone sensitivity, with artefacts appearing regardless of facial pose on darker skin. EleGANT can look visually strong on lighter skin, but it is the least reliable on darker complexions, where it introduces the most severe structural artefacts and chromatic instability.

Across-model comparison. Across models, the dominant failure mode is cross-skin-tone transfer, where foundation and global colour shifts become entangled with cosmetic style. These effects are most pronounced for Black-presenting faces, aligning with frequent tonal mismatch, chromatic noise, and regional bleeding. BeautyGAN is generally the most coherent, whereas PSGAN and EleGANT show greater instability under tone differences, limiting suitability for inclusive augmentation.

Makeup transfer (quantitative). Table 1 shows that BeautyGAN yields the lowest mean FID, while EleGANT and PSGAN induce larger overall shifts. Across all models, Black faces obtain the highest FID indicating the largest distribution shift under matched reference styles. This trend is consistent with the qualitative observations of stronger tonal mismatch and more frequent artefacts on darker skin tones. Table 2 shows the same group also exhibits the largest positive luminance shift ($\Delta L^* > 0$), suggesting systematic brightening in the central facial region compared to other groups. In contrast, Asian and White groups tend to show mild average darkening, supporting the observation that foundation transfer can modify apparent skin tone beyond localised cosmetics. Table 3 reports identity preservation under makeup transfer. Similarity remains high across models, indicating minimal identity change; BeautyGAN achieves the highest preservation, while PSGAN shows the largest cross-group variation ($\Delta=0.058$).

Hairstyle Transfer Results

We evaluate hairstyle transfer on the same bald female dataset used for makeup experiments, restricting analysis to Black subjects (40 neutral, 40 smiling). Unlike prior work that focuses on straight or wavy hair, we target underrepresented textured styles. Five reference hairstyles (afros, curls, braids, locs) were selected due to

Table 1: FID ($\downarrow$) between bald source faces ($N=80$) and makeup-transferred outputs ($N=640$ from 8 reference styles), reported per ethnicity-presenting group. Mean is averaged across groups; Δ FID is max–min.

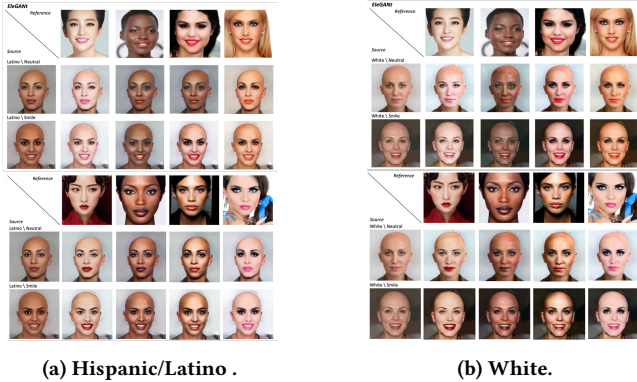
Model	Asian	Black	Hispanic/Latino	White	Mean	Δ FID
EleGANT	78.07	79.09	66.29	67.58	72.76	12.80
BeautyGAN	46.43	59.76	44.80	41.61	48.15	18.15
PSGAN	74.51	77.77	68.32	67.77	72.09	10.00

Table 2: Mid-face luminance shift under makeup transfer, reported as median $|\Delta L^*|$ (mean ΔL^* in parentheses) across ethnicity-presenting groups over $N=640$ outputs. Positive mean ΔL^* indicates brightening.

Model	Asian	Black	Hispanic/Latino	White	Mean $ \Delta L^* $	$\Delta \Delta L^* $
EleGANT	6.84 (-4.99)	20.43 (+17.80)	9.14 (+4.20)	6.15 (-2.26)	10.64	14.28
BeautyGAN	6.81 (-5.01)	13.74 (+13.74)	5.97 (+3.16)	6.44 (-3.66)	8.24	7.76
PSGAN	6.15 (-2.12)	21.28 (+19.98)	8.46 (+6.64)	6.04 (+0.39)	10.48	15.24

Table 3: ArcFace identity preservation summary ($\uparrow$) under makeup transfer. Mean and Δ are computed over group-wise median cosine similarity.

Model	Mean median cosine	Δ
EleGANT	0.781	0.024
BeautyGAN	0.892	0.012
PSGAN	0.790	0.058

**Figure 8: Hair transfer results by Barbershop [50] on bald source images. Output_01–02: afro/curly references; Output_03: braids; Output_04–05: loc references.****Figure 7: Makeup transfer results using EleGANT on Hispanic/Latino and White.**

the limited representation of these textures in existing datasets; references were sourced from high-resolution, front-facing public images to minimise pose and blur confounds. Because these styles are culturally associated, we restrict hair-transfer evaluation to Black female faces to avoid culturally inappropriate cross-group transfer and to maintain a controlled analysis. We test three GAN-based models—Barbershop and StyleYourHair (optimization-based) and HairFastGAN (encoder-based). Each identity is paired with all five references, yielding 400 outputs per model. Outputs are assessed using a fixed qualitative checklist covering hairline blending, texture fidelity, structural accuracy, and identity preservation.

Figures 8–10 show representative results; additional grids appear in the supplementary material.

Barbershop shows partial success in transferring the global structure of afro-textured hairstyles but struggles with fine-grained fidelity and integration (Figure 8). For afros (Output_01–02), the overall silhouette and volume are captured, yet curl detail is over-smoothed and hairline transitions appear abrupt. In braided styles (Output_03), directional flow and parting are reproduced, but geometric regularity and strand separation are poorly rendered, with braids appearing fused or flattened. Loc-based references (Output_04–05) present greater difficulty: loc density and cylindrical structure are misrepresented, often resembling twisted strands rather than realistic locs. Scalp-level artefacts, including clipping and root misalignment, further reduce realism. Although facial identity is preserved, hairstyle quality deteriorates with increasing structural complexity, highlighting limitations in modelling dense, high-frequency textures.

StyleYourHair shows partial success in transferring afro-textured hairstyles but struggles with texture fidelity and scalp blending (Figure 9). Although global shape is generally preserved across afros, braids, and locs, high-frequency detail is oversmoothed. Afro styles (Output_01–02) retain volume but lack curl definition, while braids and locs (Output_03–05) lose strand separation and cylindrical structure, often appearing fused or flattened. Scalp integration is inconsistent, with visible seams in some cases. Facial identity remains

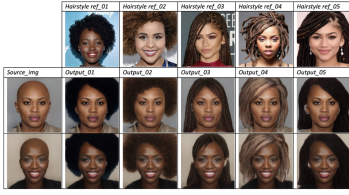


Figure 9: Hair transfer results by StyleYourHair [27] on bald source images.



Figure 10: Hair transfer results by HairFastGAN [35] on bald source images.

stable, yet realism declines with increasing hairstyle complexity, revealing limitations in modelling dense, textured styles.

HairFastGAN demonstrates relatively stronger performance on afro-textured hairstyles compared to the other models, particularly in afro and curly styles (Output_01 and Output_02; Figure 10). The model effectively captures the global shape and volume of afros, with improved texture fidelity and natural facial blending. Side parts and asymmetries are preserved more faithfully than in the other models. However, finer coil details are still smoothed out, resulting in broader, less defined textures that limit photorealism. For braided (Output_03) and loc styles (Output_04 and Output_05), performance declines. Braids appear merged or oversimplified, lacking the geometric regularity of real strands. Locs lose their rope-like definition, resembling loosely grouped layers. While identity retention remains strong, hair colour consistency is less reliable, with some styles showing noticeable shifts in tone. HairFastGAN aligns shape and volume well but lacks strand-level detail for high-fidelity transfer of textured styles.

Hairstyle transfer (quantitative). Table 4 reports LPIPS [47] between each fixed reference hairstyle and its corresponding generated outputs (lower is better). Across models, Ref 01 consistently yields the highest distances (0.589–0.699), indicating greater perceptual deviation from the target style relative to other references. In contrast, Ref 02 and Ref 05 exhibit lower distances across models. Overall, StyleYourHair achieves the lowest mean LPIPS (0.484), followed by Barbershop (0.494), while HairFastGAN shows the highest overall distance (0.527). Performance differences are reference-dependent, suggesting that hairstyle transfer fidelity varies substantially with the structural and textural characteristics of the target style.

Discussion

A primary source of bias in generative models lies in the composition of their training data. When certain skin tones are under-represented, learned representations become skewed, leading to

Table 4: Perceptual similarity LPIPS ($\downarrow$) between each reference hairstyle and generated output for Black-presenting inputs. Values are mean LPIPS averaged over neutral and happy expressions ($N=80$ outputs per reference; $N=400$ per model). Lower is better. Best per column is bold.

Model	Ref 01	Ref 02	Ref 03	Ref 04	Ref 05	Mean
Barbershop	0.664	0.407	0.571	0.455	0.371	0.494
StyleYourHair	0.589	0.320	0.492	0.555	0.463	0.484
HairFastGAN	0.699	0.441	0.491	0.533	0.470	0.527

systematic performance gaps. This pattern is well documented in computer vision, where models trained predominantly on lighter-skinned faces perform poorly on darker skin tones, and it directly extends to makeup transfer systems. Widely used datasets such as the Makeup Transfer (MT) dataset [29] are dominated by East Asian and Caucasian faces, with limited representation of darker skin tones.

Our qualitative and quantitative results reflect these limitations. Tonal shifts, ashy overlays, lipstick misalignment, chromatic noise indicate failures in colour adaptation and weak disentanglement between cosmetic style and skin tone. When GANs lack sufficient examples of dark-skinned individuals wearing diverse makeup styles, they struggle to generalise across skin tones. Prior work shows that demographic labels and skin tone annotations are often subjective and inconsistent [3, 26], further distorting model learning. Related studies also report that image-to-image makeup models frequently alter skin tone alongside cosmetic features, highlighting weak separation between content and style [7, 33]. Evaluation practices further compound these issues. Performance is often demonstrated using unrepresentative samples or visually favourable outputs, creating an inflated perception of model reliability. Without balanced datasets and rigorous evaluation protocols, GAN-based beauty applications risk reinforcing existing disparities rather than mitigating them. Hairstyle transfer models exhibit analogous limitations. Although Barbershop [50], StyleYourHair [27], and HairFastGAN [35] perform reasonably on straight and loosely wavy hairstyles, fidelity decreases for afro-textured and protective styles (e.g., braids and locs). This pattern aligns with prior observations that many hairstyle datasets remain skewed toward globally smooth hair geometries and lack detailed annotation of complex Black hairstyles [8, 41]. Consistent with these qualitative trends, LPIPS results in Table 4 show increased perceptual distance for structurally complex references, indicating reduced texture and hairline fidelity across models. Optimisation-based approaches capture coarse global shape but struggle with strand separation and scalp integration [9], while encoder-based methods such as HairFastGAN improve blending yet still exhibit higher perceptual deviation on dense, coiled textures.

These limitations extend beyond technical concerns. Hair texture and style are closely tied to identity and cultural expression, as recognised by legal protections such as the CROWN Act [1, 2]. Inadequate synthesis of textured hair risks aesthetic erasure and algorithmic discrimination, particularly in public-facing and virtual environments. More broadly, makeup and hairstyle transfer systems

can encode and reproduce dominant beauty norms. Consistently stronger performance on lighter skin tones and straighter hairstyles risks reinforcing colourist biases and excluding marginalised appearance groups from reliable digital representation. Mitigating these effects requires diverse training data, improved disentanglement of style and identity, and evaluation protocols that explicitly account for demographic equity.

Conclusions

Our findings show that existing models systematically struggle with cross-skin tone makeup transfer and with textured, culturally specific hair types, introducing artefacts, tonal distortions, and structural failures. These issues are most pronounced for Black-presenting faces, where the combination of darker skin tones and afro-textured/protective hairstyles exposes blind spots in prevailing training data and incomplete disentanglement between style attributes (makeup/hair) and identity-related appearance. As such tools are increasingly used in public-facing applications and augmentation pipelines, demographic inconsistencies raise both technical concerns and risks of exclusion or cultural misrepresentation. We recommend more diverse and better-annotated training datasets, architecture-level improvements that explicitly separate content from style, and evaluation protocols that account for equity across skin tones and hair textures. Future work will expand quantitative fairness and image-quality evaluation, broaden demographic coverage to include additional ethnicities and male subjects, and explore alternative transfer frameworks, including improved GAN-based approaches [37, 48] and diffusion-based editing models [4, 36].

References

- [1] 2019. CROWN Act. <https://www.thecrownact.com/>. Accessed: 2025-04-29.
- [2] 2019. NYC Human Rights Law - Legal Guidance on Hair Discrimination. <https://www.nyc.gov/site/cchr/law/hair-discrimination-legal-guidance.page>. Accessed: 2025-04-29.
- [3] Teanna Barrett, Quanzen Chen, and Amy Zhang. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1757–1771. doi:10.1145/3593013.3594114
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.
- [5] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. 2021. Adaptive Convolutions for Structure-Aware Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7972–7981.
- [6] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. 2018. PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 40–48. doi:10.1109/CVPR.2018.00012
- [7] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. 2019. BeautyGlow: On-Demand Makeup Transfer Framework With Reversible Generative Network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10034–10042. doi:10.1109/CVPR.2019.01028
- [8] Yutao Chen, Yuxuan Zhang, Zhongrui Huang, Zhenyao Luo, and Jinpeng Chen. 2021. Celebhair: A new large-scale dataset for hairstyle recommendation based on celeba. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 323–336.
- [9] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. 2020. Editing in Style: Uncovering the Local Semantics of GANs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5770–5779. doi:10.1109/CVPR42600.2020.00581
- [10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [11] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 289–293.
- [12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.
- [13] Generated Media, Inc. 2024. Generated Photos: 100,000 AI-Generated Faces. <https://generated.photos/>. Accessed: 2024-11-02.
- [14] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. 2021. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision* 129, 7 (2021), 2288–2307.
- [15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) (NIPS'14). MIT Press, Cambridge, MA, USA, 2672–2680.
- [16] Alvin Grissom II, Ryan F. Lei, Matt Gusdorff, Jeova Farias Sales Rocha Neto, Bailey Lin, and Ryan Trotter. 2024. Examining Pathological Bias in a Generative Adversarial Network Discriminator: A Case Study on a StyleGAN3 Model. *arXiv preprint arXiv:2402.09786* (2024). <https://arxiv.org/abs/2402.09786>
- [17] Alex Hernández-García and Peter König. 2018. Further Advantages of Data Augmentation on Convolutional Neural Networks. In *Proceedings of the 27th International Conference on Artificial Neural Networks (ICANN 2018), Part I (Lecture Notes in Computer Science, Vol. 11139)*. Springer, 95–103.
- [18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6629–6640.
- [19] Minui Hong, Jinwoo Choi, and Gunhee Kim. 2021. StyleMix: Separating Content and Style for Enhanced Data Augmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 14857–14865. <https://api.semanticscholar.org/CorpusID:235703442>
- [20] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1510–1519. doi:10.1109/ICCV.2017.167
- [21] Philip T. Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P. Breckon, and Boguslaw Obara. 2019. Style Augmentation: Data Augmentation via Style Randomization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 83–92.
- [22] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. 2022. Imperfect ImaGANation: Implications of GANs exacerbating biases on facial data augmentation and snapchat face lenses. *Artificial Intelligence* 304 (2022), 103652. doi:10.1016/j.artint.2021.103652
- [23] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. 2019. PSGAN: Pose and Expression Robust Spatial-Aware GAN for Customizable Makeup Transfer. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 5193–5201. <https://api.semanticscholar.org/CorpusID:214181220>
- [24] Youngjoo Jo and Jongyoul Park. 2019. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1745–1753. doi:10.1109/ICCV.2019.00183
- [25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Hk99zCeAb>
- [26] Zaid Khan and Yun Fu. 2021. One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 587–597. doi:10.1145/3442188.3445920
- [27] Taewoo Kim, Chaeyeon Chung, Yoonseo Kim, Sunghyun Park, Kangyeol Kim, and Jaegul Choo. 2022. Style Your Hair: Latent Optimization for Pose-Invariant Hairstyle Transfer via Local-Style-Aware Hair Alignment. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer-Verlag, Berlin, Heidelberg, 188–203. doi:10.1007/978-3-031-19790-1_12
- [28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. doi:10.1145/3065386
- [29] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*. 645–653.
- [30] Swee Kiat Lim, Yi Loo, Ngoc-Trung Tran, Ngai-Man Cheung, Gemma Roig, and Yuval Elovici. 2018. DOPING: Generative Data Augmentation for Unsupervised

- Anomaly Detection with GAN. *2018 IEEE International Conference on Data Mining (ICDM)* (2018), 1122–1127. <https://api.semanticscholar.org/CorpusID:52075217>
- [31] Yun Luo and Bao-Liang Lu. 2018. EEG Data Augmentation for Emotion Recognition Using a Conditional Wasserstein GAN. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2535–2538. doi:10.1109/EMBC.2018.8512865
- [32] Vongani H. Mahuleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A. Efros, Angjoo Kanazawa, and Devin Guillory. 2022. Studying bias in gans through the lens of race. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII* (Tel Aviv, Israel). Springer, Berlin, Heidelberg, 344–360. doi:10.1007/978-3-031-19778-9_20
- [33] Masoud Mokhtari, Fatemeh Taheri Dezaki, Timo Bolkart, Betty Mohler Tesch, Rahul Suresh, and Amin Banitalebi-Dehkordi. 2024. Skin Tone Disentanglement in 2D Makeup Transfer With Graph Neural Networks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2745–2749. doi:10.1109/ICASSP48485.2024.10446354
- [34] Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array* 16 (2022), 100258. doi:10.1016/j.array.2022.100258
- [35] Maxim Nikolaev, Mikhail Kuznetsov, Dmitry P. Vetrov, and Aibek Alanov. 2024. HairFastGAN: Realistic and Robust Hair Transfer with a Fast Encoder-Based Approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (Eds.), Vol. 37. Curran Associates, Inc., 45600–45635. https://proceedings.neurips.cc/paper_files/paper/2024/file/50f1dd93c57ba57fe246fba9e64f5b20-Paper-Conference.pdf
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. doi:10.1109/CVPR52688.2022.01042
- [37] Rohit Saha, Brendan Duke, Florian Shkurti, Graham W. Taylor, and Parham Aarabi. 2021. LOHO: Latent Optimization of Hairstyles via Orthogonalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1984–1993.
- [38] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific reports* 9, 1 (2019), 16884.
- [39] Sefik Serengil and Alper Ozpinar. 2024. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Journal of Information Technologies* 17, 2 (2024), 95–107. doi:10.17671/gazibtd.1399077
- [40] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). <https://api.semanticscholar.org/CorpusID:14124313>
- [41] Haiyu Wu, Grace Bezold, Aman Bhatta, and Kevin W Bowyer. 2023. Logical consistency and greater descriptive power for facial hair attribute learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8588–8597.
- [42] Yiqian Wu, Yong-Liang Yang, and Xiaogang Jin. 2022. HairMapper: Removing hair from portraits using gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4227–4236. doi:10.1109/CVPR52688.2022.00419
- [43] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. 2022. EleGANt: Exquisite and Locally Editable GAN for Makeup Transfer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 737–754. doi:10.1007/978-3-031-19787-1_42
- [44] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseok Joo. 2022. Enhancing Fairness in Face Detection in Computer Vision Systems by Demographic Bias Mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) (*AIES '22*). Association for Computing Machinery, New York, NY, USA, 813–822. doi:10.1145/3514094.3534153
- [45] Wei Yi, Yaoran Sun, and Sailing He. 2018. Data Augmentation Using Conditional GANs for Facial Emotion Recognition. In *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)*. 710–714. doi:10.23919/PIERS.2018.8598226
- [46] Seyma Yucer, Samet Akcay, Noura Al-Moubayed, and Toby P. Breckon. 2020. Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, Los Alamitos, CA, USA, 83–92. doi:10.1109/CVPRW50498.2020.00017
- [47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
- [48] Huihuang Zhao, Lin Zhang, Paul L. Rosin, Yu-Kun Lai, and Yaonan Wang. 2024. HairManip: High quality hair manipulation via hair element disentangling. *Pattern Recognition* 147 (2024), 110132. doi:10.1016/j.patcog.2023.110132
- [49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.
- [50] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. 2021. Barbershop: GAN-based image compositing using segmentation masks. *ACM Transactions on Graphics* 40, 6 (Dec. 2021), 1–13. doi:10.1145/3478513.3480537
- [51] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. 2018. Emotion Classification with Data Augmentation Using Generative Adversarial Networks. In *Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 10939)*, Dinh Q. Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi (Eds.). Springer, 349–360. doi:10.1007/978-3-319-93040-4_28