DIFFCLEAN: Diffusion-based Makeup Removal for Accurate Age Estimation

Ekta Gavas
New York University
eg4131@nyu.edu

Sudipta Banerjee
University of Wyoming
sbanerj3@uwyo.edu

Chinmay Hegde
New York University
chinmay.h@nyu.edu

Nasir Memon
New York University
memon@nyu.edu

1. Failure Case Analysis

Fig. 1 shows examples of failure cases of makeup removal by DIFFCLEAN on the LADN dataset. Failure cases include drastic Halloween-style makeup, where our method can reduce the effect of makeup traces but not completely remove them. **In the future, will use semantic segmentation masks to drive locally adaptive makeup removal against extreme makeup.**

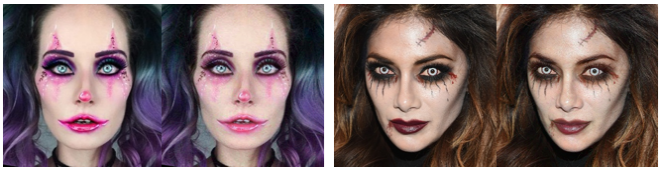


Figure 1. Failure cases of DIFFCLEAN on extreme Halloween-style makeup images.

2. Additional evaluation

We present additional analysis by evaluating on new synthetic makeup data (IMDB-Clean), more results on baseline (MAD) and new face recognition model (IR-152).

(A) Data. We test our method on a synthetic makeup dataset *IMDB-Clean* [3], which is the age annotated version of IMDB-Clean. We used EleGANt [7] to curate the synthetic makeup transfer. We observe the following results. In terms of Minor/Adult accuracy(↑): 93.7% (Makeup), 87.3% (CLIP2Protect), **94.6**% (Ours-SSRNet). In terms of MAE(↓): 2.5 (Makeup), 3.4 (CLIP2Protect), **2.5** (Ours-SSRNet). Our method results in higher accuracy and lower MAE. We provide additional visual comparisons on example images from CelebA-HQ dataset in Fig. 2

(B) Baseline. We present more examples of comparison between, DIFFCLEAN and MAD [4]; see Fig. 3. MAD struggles to restore original age and may introduce visible artifacts. The artifacts are more evident if the test dataset (CelebA-HQ) is different than the training dataset (MT) showing limited generalizability. In contrast, our method can handle cross-dataset and cross-style makeup removal. Even when tested on trained data (MT), it introduces hallucinations such as lower teeth in Fig. 3 (1st row), which are not present in the original image.

Table 1. Performance on BeautyFace dataset using IR-152 and MobileFace matchers in terms of genuine similarity scores(↑).

Method	IR-152	MobileFace
CLIP2Protect	0.85	0.83
DiffClean (Ours-SSRNet)	0.92	0.95
DiffClean (Ours-CLIP)	0.91	0.94

(C) Face recognition. We computed the cosine similarity scores for the same identity (only genuine pairs) before makeup removal and after makeup removal using our method, DIFFCLEAN on the BeautyFace dataset using two different face matchers: IR-152 and MobileFace. We further compared it with the results of CLIP2Protect. The purpose of this experiment is to examine if our method introduces artifacts that inadvertently lowers the biometric similarity, resulting in false non-matches. As seen from the results in Table 1, our method produces higher genuine similarity compared to CLIP2Protect, thereby having lesser chances of false non-matches.

3. Real-world makeup data analysis

We present the performance of our method in terms of identity verification on LADN dataset [1] in Fig. 4 and image quality evaluation on LADN and Makeup-Wild datasets in Table 2.

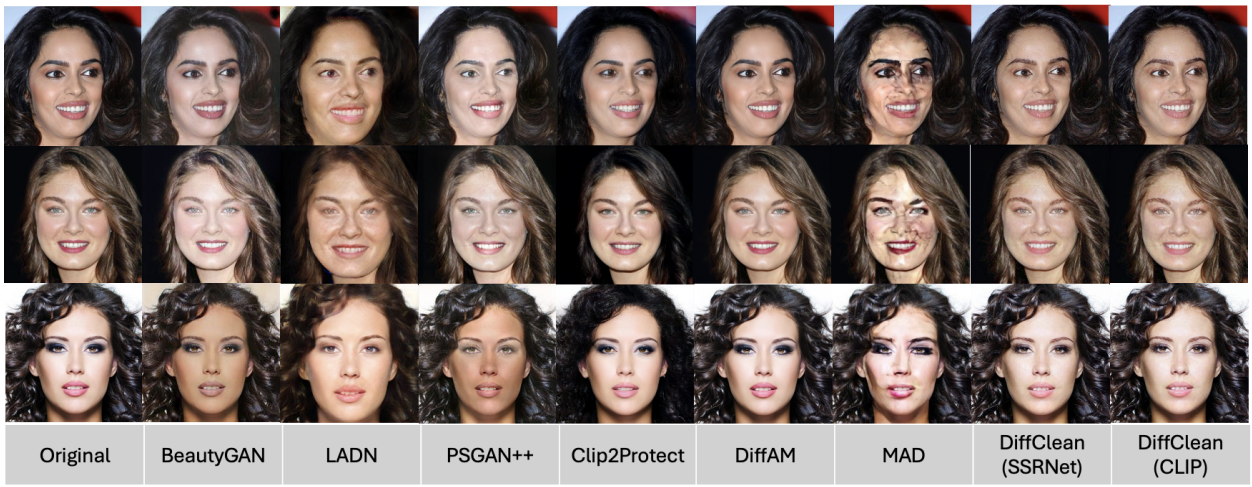


Figure 2. Comparison of makeup removal results generated by six baselines and our proposed DIFFCLEAN (last two columns) on three example images from CelebA-HQ [2] dataset. GAN-based baselines (BeautyGAN, LADN, PSGAN++) introduce visual artifacts, while CLIP2Protect alters hair color and style, DiffAM does not effectively remove makeup, and MAD produces distortions on unseen data.

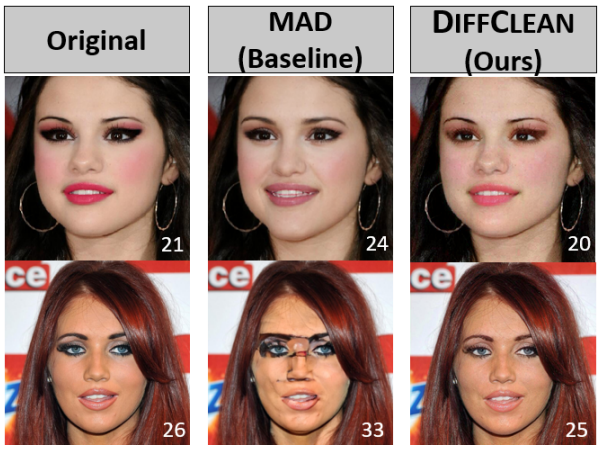


Figure 3. Results of makeup removal on real-world makeup images from MT (top row) and CelebA-HQ (bottom row) datasets by MAD (baseline) and DIFFCLEAN (our method). Note our method is capable of restoring the correct age (indicated in white) compared to MAD.

Table 2. Results of makeup removal in terms of image quality metrics (SSIM↑ and PSNR↑) on real-world makeup datasets LADN and Makeup-Wild.

Method	LADN		Makeup-Wild	
	SSIM	PSNR	SSIM	PSNR
Ours-DIFFCLEAN (SSRNet)	0.94	30.77	0.97	35.09
Ours-DIFFCLEAN (CLIP)	0.94	30.59	0.97	34.97

4. Performance breakdown by age groups

We present a detailed performance breakdown of DIFFCLEAN on each age group in Table 3. Note that CLIP-based age loss is marginally better than SSRNet-based age loss. Our method improves the age estimation accuracy on the *target age* groups (minors/teenagers): [10-14] from 25% (CLIP2Protect) and 28% (DiffAM) to **46%** (Ours); age group [15-19] from 26% (CLIP2Protect) and 31% (DiffAM) to **43%** (Ours) as seen in Table 3, while sacrificing accuracy in [20-29] age group by 8%. On average, our method achieves the lowest mean and standard deviation of MAE across 9 age groups, as follows: 6.53 ± 2.4 (CLIP2Protect), 6.31 ± 2.1 (DiffAM), 5.75 ± 1.7 (Ours-SSRNet) and $\mathbf{5.70} \pm \mathbf{1.7}$ (Ours-CLIP).

We further investigated the drop in performance in the [20-29] group by comparing the ground-truth age with predicted age on original vs. makeup removed images in FFHQ. We observed MAE: 3.75 vs. 4.11, number of underestimation errors: 24 vs. 51, number of overestimation errors: 64 vs. 45. This shows that our method lowers the overestimation errors in predicted age due to makeup at the expense of overall higher MAE.

5. Importance of Weighted Self-Adjusted Smoothed L1 Loss Function

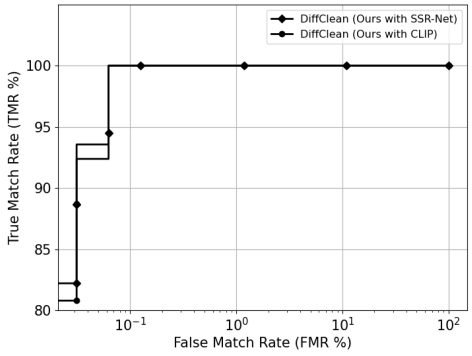
Eqn.(1) in the main paper refers to the self-adjusted smoothed $\mathcal{L}_1$ loss function adopted in RetinaMask which is weighted by 3.0 for the vulnerable age groups between 10-29 yrs. It produced the lowest MAE compared to other losses; see 4.

6. Impact on adversarial makeup transfer-based privacy protection

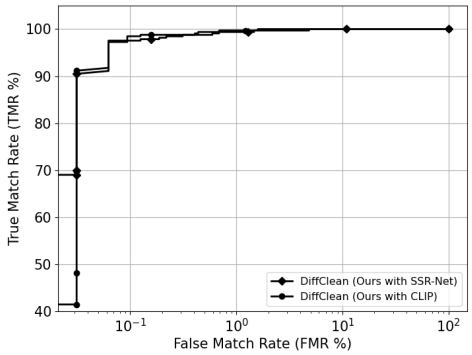
Both CLIP2Protect and DiffAM are essentially makeup transfer-based privacy protection schemes that imperson-

Table 3. Results of age estimation accuracy (%) ↑ on each age group on the FFHQ dataset.

Age group	Makeup images	CLIP2Protect ('no makeup') prompt	DiffAM (MR)	DIFFCLEAN (Ours) SSRNet age loss	DIFFCLEAN (Ours) Clip age loss
0-2	0.0	0.0	0.0	0.0	0.0
3-6	2.1	1.8	2.8	1.8	1.4
7-9	3.5	2.5	3.9	3.9	5.0
10-14	26.5	25.1	28.6	45.5	46.2
15-19	28.8	26.8	30.6	43.4	43.0
20-29	75.6	72.6	68.0	65.0	65.0
30-39	50.1	47.1	48.0	48.7	46.9
40-49	53.7	58.1	58.1	53.4	55.5
50-69	70.2	66.3	72.2	70.7	70.7



FaceNet



MobileFace

Method	FR model	
	FaceNet	MobileFace
Ours-DIFFCLEAN (SSRNet)	**81.6**	**69.0**
Ours-DIFFCLEAN (CLIP)	80.2	41.5

Figure 4. (Top): ROC curve with FaceNet. (Middle): ROC curve with MobileFace. (Bottom): Biometric matching in terms of TMR (%) @FMR = 0.01% (*higher is better*) with FaceNet and MobileFace matchers on LADN dataset.

ate a targeted identity to fool the face matcher. DIFFCLEAN removes makeup using a diffusion model, and we wanted to investigate whether it affects the attack success rate (ASR) [6] of the protected faces on MobileFace (threshold=0.302) at False Match Rate=0.01 following [5]. $ASR = \frac{\#sim(protectedface, targetid) > th}{\#no.of comparisons}$. We conducted a preliminary analysis using 1,000 images from CelebA-HQ [2] dataset and one targeted identity for impersonation attack with 1 prompt for CLIP2Protect ('Matte') and 1 reference style from DiffAM ('XMY-060'). Our initial findings indicate that DIFFCLEAN is benign and does not compromise the privacy of protected faces (see Table 5). In the future, we will comprehensively analyze this impact on various protection mechanisms across multiple makeup styles.

Table 4. Comparison of $\mathcal{L}_{WSL}$ with $\mathcal{L}_1$, $\mathcal{L}_2$, and Huber loss functions for age estimator.

MAE with different age losses					
Age Group	**Pretrained**	**L1 Loss**	**MSE Loss**	**Huber Loss**	$\mathcal{L}_{WSL}$ **(Eqn.1)**
0-2	20.2	2.1	3.1	2.6	1.6
3-6	27.9	2.3	2.6	2.5	2.2
7-9	29.2	3.7	4.1	3.7	3.1
10-14	26.4	5.0	5.1	5.1	4.6
15-19	18.8	5.6	6.1	6.3	5.3
20-29	14.6	6.4	6.2	6.5	6.1
30-39	12.6	6.6	6.0	6.4	6.0
40-49	9.7	8.0	7.0	7.7	6.8
50-69	9.4	9.7	8.7	8.9	6.9
Average	18.7	5.5	5.4	5.5	**4.7**

Table 5. Attack Success Rate (ASR) (↑) with MobileFace after makeup removal with DIFFCLEAN on protected faces generated using CLIP2Protect and DiffAM.

Methods	**Attack Success Rate (ASR)**	
	CLIP2Protect	DiffAM
Protected Faces	83.04	58.62
DIFFCLEAN (Ours-SSRNet)	84.35 (↑)	61.45 (↑)
DIFFCLEAN (Ours-CLIP)	84.56 (↑)	62.80 (↑)

References

[1] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. LADN: Local adversarial disentangling net-

work for facial makeup and de-makeup. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10481–10490, 2019. 1

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 3

[3] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. FP-age: Leveraging face parsing attention for facial age estimation in the wild. *IEEE Transactions on Image Processing*, 2022. 1

[4] Bo-Kai Ruan and Hong-Han Shuai. MAD: Makeup All-in-One with Cross-Domain Diffusion Model. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 1

[5] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20595–20605, 2023. 3

[6] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24584–24594, 2024. 3

[7] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. Elegant: Exquisite and locally editable gan for makeup transfer. In *European Conference on Computer Vision*, pages 737–754. Springer, 2022. 1