

# Appendix

## A Details of Our Benchmark

First, as illustrated in Fig. S1, we present our criteria for categorizing our proposed VintageFace benchmark into simple, medium, and severe degradation levels. Specifically, we employ a frozen CLIP model to compute the similarity between each old photo and a textual description of degradation severity. Images are then ranked by their similarity scores and assigned to categories accordingly. To ensure accuracy, we further manually corrected a small number of misclassified samples.

Second, as shown in Fig.S2, we display representative examples from each degradation level in the VintageFace benchmark. These examples demonstrate varying degrees of blurring, fading, and structural damage, and are largely consistent with the classification criteria established in Fig.S1.

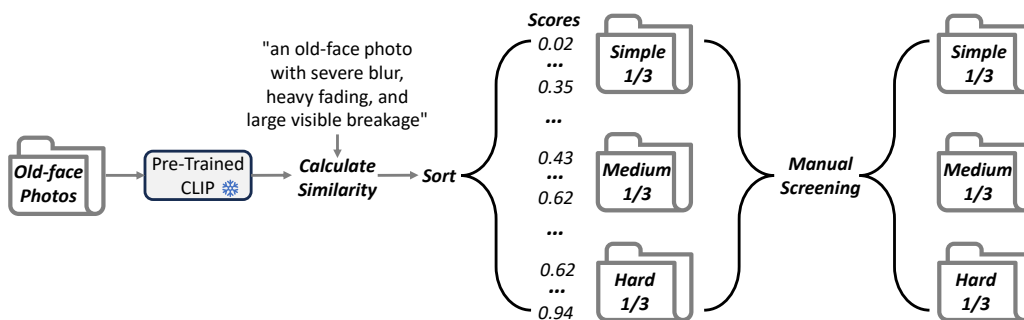


Figure S1: Method for categorizing degradation types into simple, medium, and hard levels in our benchmark VintageFace for testing.



Figure S2: A showcase of representative facial images with varying degradation types from our VintageFace benchmark. The benchmark includes faces across diverse genders, ages, and ethnicities.

## B More Comparisons

First, we provide additional visual results in the appendix (*e.g.*, Fig.S5, Fig.S6, and Fig.S7) to complement the main text. These figures showcase the restoration performance on old face photographs

Table S1: Quantitative comparison on real-world BFR benchmarks that contain old face photos, like WebPhoto-Test and CelebA-Child. **Bold** and underlined indicate best and second best results.

Dataset	Metric	GAN-based		Diffusion-based (Learning)		Diffusion-based (Train-free)		
		GPEN [1]	Code [2]	DiffFace [3]	DiffBIR [4]	DDNM [5]	PGDiff [6]	Ours
<b>WebPhoto-Test</b>	FID↓	101.3	<b>83.2</b>	89.1	91.8	165.6	96.1	<u>86.9</u>
	NIQE↓	6.326	4.705	4.831	6.069	9.259	5.117	<b>4.406</b>
<b>CelebA-Child</b>	FID↓	113.0	116.2	113.1	118.9	151.2	121.0	<b>112.7</b>
	NIQE↓	4.945	4.983	<u>4.818</u>	5.549	6.576	5.070	<b>4.524</b>

across diverse genders, ethnicities (Asian, European-American, and Indian), and age groups. Our method effectively balances perceptual quality and identity preservation: the restored images exhibit minimal artifacts or breakage while maintaining faithful facial identity.

Second, we observe that widely used blind face restoration benchmarks, such as LFW-Test and CelebChild, also include a substantial number of old face photos. However, these differ from our dataset in that they primarily exhibit blurring, with no significant structural damage and limited fading. To demonstrate the generalization and effectiveness of our method, we compare it with state-of-the-art approaches on these two benchmarks. Following previous works [2, 6], we adopt FID and NIQE as evaluation metrics. As shown in Table S1, our method achieves good quantitative results on both benchmarks. Furthermore, we provide visual comparisons in Fig. S8, which reveal that our method not only effectively addresses blurring but also excels at restoring facial color. This perceptual advantage, particularly in color restoration, is not fully captured by quantitative metrics.

Thirdly, VintageFace primarily consists of frontal photos, as portrait photos decades ago were typically studio-based, focusing on clearly capturing facial features, making profile shots rare. Additionally, eyewear was less common, resulting in fewer photos with glasses. Consequently, our data has fewer such samples. Nevertheless, as shown in Fig. S3, SSDiff performs robustly across these scenarios, including glasses, profile shots, and severe degradations, consistently yielding favorable results.

## C More Ablation

**Robustness of Pre-trained Networks.** Our SSDiff is generally robust to inaccuracies in external components (face parsing, scratch detection, style transfer). These networks only provide coarse directional signals during reverse diffusion, similar to classifier-guided diffusion, and are not strict constraints. As long as the guidance is not severely misleading, the strong generative prior of the frozen diffusion model dominates reconstruction. To quantify this robustness, as shown in the Table S2, Table S3, and Table S4, we conduct ablations on the Medium type subset:

For parsing map networks, we introduce inaccuracies by replacing the original parsing maps with pseudo-label parsing maps of different strengths  $s$ , where the resulting errors are even larger than those observed in parsing maps under severe degradations (79% IOU). The resulting IoU with the original parsing map is: for  $s=2.5e-4$ , IoU=76% (24% discrepancy); for  $s=1e-4$ , IoU=70% (30% discrepancy); For scratch detection networks, we randomly flip 10%, 20%, and 30% of the masks of breakage regions to simulate a situation where some of the breakages have not been detected; For style transfer networks, we weaken the style transfer guidance by reducing the style factor  $\alpha$  from 0 to 0.1 and 0.2, slightly affecting color and content.

Table S2: Parsing Maps.				Table S3: Scratch Masks.				Table S4: Style Transfer.		
	Ours	76%	70%	Ours	10%flip	20%flip	30%flip	Ours ( $\alpha=0$ )	$\alpha=0.1$	$\alpha=0.2$
FID(↓)	128.3	131.1	133.4	128.3	129.2	130.7	132.5	128.3	129.2	128.8
MAN-IQA(↑)	0.395	0.391	0.382	0.395	0.391	0.379	0.381	0.395	0.396	0.392
Face Sim.(↑)	1	0.985	0.955	1	0.974	0.953	0.937	1	0.977	0.959

Here, Face Similarity (range [0, 1]) denotes the cosine similarity between features (extracted with ArcFace [7]) of the perturbed restoration and the original restoration (Ours). These results show that errors in the pre-trained networks are not severe, and the strong generative prior of the diffusion model can propagate the correct cues to other regions, preventing significant performance drops. This demonstrates that SSDiff is robust to these pre-trained networks.

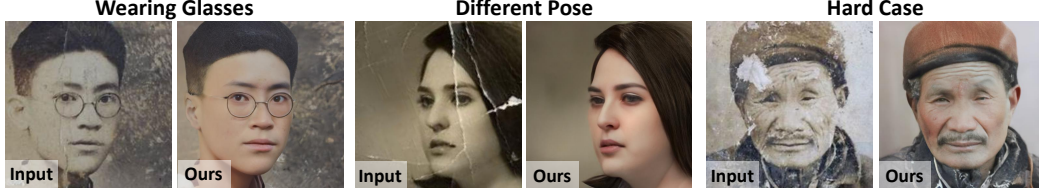


Figure S3: Visualization of SSDiff under wearing glasses, different poses, and severe degradation.

**Latency.** Our method is built upon existing pre-trained diffusion face generation models, where additional inference overhead mainly comes from four components: a simple restore, a face parsing network, a scratch detection network, and a style transfer network. All components are lightweight and are executed at a single denoising step rather than throughout the entire process. Moreover, except for the style migration network, the other three can optionally be pre-processed offline. When all four components are executed online, the average latency for processing a single old face photo is 95 ms on an NVIDIA GeForce RTX 4090. If the three offline-optional components are pre-processed, the average latency is reduced to 24 ms. In contrast, PGDiff [6] requires semantic information extraction at each denoising step, introducing a latency of about 10 s. Therefore, our method only introduces minimal latency to existing pre-trained generation diffusion frameworks.

**Computational Cost.** As shown in Table S5, we further compare the number of Params, FLOPs, and inference time of our method with existing diffusion-based face restoration methods [3, 4, 6]. We let the restore be performed offline; our method performs excellently. Our method is smaller in terms of the Params and FLOPs counts, especially compared to stable diffusion-based methods like DiffBIR [4].

Table S5: Quantitative comparison on computational costs with existing diffusion-based BFR methods.

Costs	DiffFace [3]	DiffBIR [4]	PGDiff [6]	Ours
Params↓	175.4M	1717M	47.7M	<b>45.4M</b>
FLOPs↓	268.8G	24234G	127.5G	<b>120.7G</b>

## D Broader Impacts

This work focuses on accurately restoring old face photographs, with potential applications in cultural heritage preservation and family history archiving. However, we acknowledge that the proposed method could be misused for malicious purposes, such as data forgery or identity-related fraud. We advise the public and downstream users to exercise caution and consider appropriate safeguards when deploying such technologies.

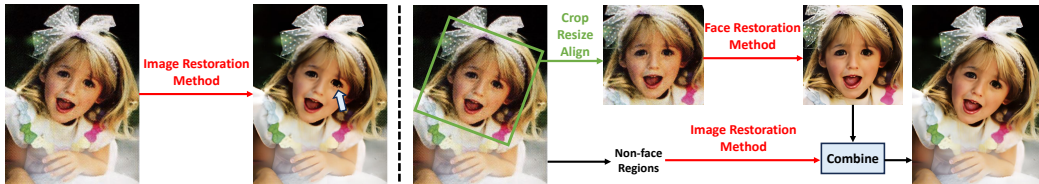


Figure S4: **(Left)** Face images are highly sensitive to artifacts, directly restoring photos containing faces with image restoration methods may result in visually disturbing results. **(Right)** A common strategy involves cropping and aligning facial regions, followed by restoration using face restoration methods, while non-facial regions are enhanced with image restoration methods to ensure visual perception. *Therefore, old photo face restoration holds practical value for old photo restoration.*

## E Necessity of Old-Photo Face Restoration

While general image restoration methods [8] aim to restore the entire image holistically, we argue that dedicated face restoration [9, 10, 11] is necessary and beneficial, especially in the context of severely degraded old portraits. As illustrated in Fig. S4, directly applying general real-world image restoration models [12] to facial regions may introduce noticeable artifacts, even when these methods perform reasonably well on background areas. This is because facial regions are typically small in size,

81 contain rich structural priors (*e.g.*, eyes, nose, mouth), and are highly sensitive to local distortions.  
 82 Artifacts in these regions are particularly perceptible and detrimental to human perception.

83 Similarly, old face photos suffer from unique degradation patterns such as heavy blurring, fading, and  
 84 structural damage. Applying global restoration methods [12, 13, 14] to these faces without region-  
 85 specific modeling frequently leads to distorted identity features or unnatural textures. Therefore,  
 86 we advocate for face-specific old photo restoration approaches [5, 6, 15] that focus on preserving  
 87 facial identity and fidelity, while allowing general old photo restoration techniques [16] to handle  
 88 the surrounding non-facial regions. This targeted strategy ensures high-quality restoration where  
 89 perceptual sensitivity is highest and complements broader restoration pipelines. Therefore, we  
 90 respectfully believe the task of old photo face restoration holds specifically practical value.

## 91 References

- 92 [1] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face  
 93 restoration in the wild. In *CVPR*, pages 672–681, 2021.
- 94 [2] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration  
 95 with codebook lookup transformer. In *NeurIPs*, volume 35, pages 30599–30611, 2022.
- 96 [3] Zongsheng Yue and Chen Change Loy. Diffface: Blind face restoration with diffused error contraction.  
 97 *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- 98 [4] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and  
 99 Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pages  
 100 430–448. Springer, 2024.
- 101 [5] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space  
 102 model. In *ICLR*, 2023.
- 103 [6] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for  
 104 versatile face restoration via partial guidance. In *NeurIPs*, 2023.
- 105 [7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for  
 106 deep face recognition. In *CVPR*, pages 4690–4699, 2019.
- 107 [8] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse  
 108 weather conditions via histogram transformer. In *ECCV*, pages 111–129. Springer, 2024.
- 109 [9] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with  
 110 generative facial prior. In *CVPR*, pages 9168–9178, 2021.
- 111 [10] Wenjie Li, Mei Wang, Kai Zhang, Juncheng Li, Xiaoming Li, Yuhang Zhang, Guangwei Gao, Weihong  
 112 Deng, and Chia-Wen Lin. Survey on deep face restoration: From non-blind to blind and beyond. *arXiv*  
 113 *preprint arXiv:2309.15490*, 2023.
- 114 [11] Wenjie Li, Heng Guo, Xuannan Liu, Kongming Liang, Jiani Hu, Zhanyu Ma, and Jun Guo. Efficient face  
 115 super-resolution via wavelet-based feature enhancement network. In *ACM MM*, pages 4515–4523, 2024.
- 116 [12] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-  
 117 resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021.
- 118 [13] Wenjie Li, Heng Guo, Yuefeng Hou, Guangwei Gao, and Zhanyu Ma. Dual-domain modulation network  
 119 for lightweight image super-resolution. *IEEE Transactions on Multimedia*, 2025.
- 120 [14] Wenjie Li, Heng Guo, Yuefeng Hou, and Zhanyu Ma. Fourierstr: A fourier token-based plugin for efficient  
 121 image super-resolution. *arXiv preprint arXiv:2503.10043*, 2025.
- 122 [15] Wenjie Li, Xiangyi Wang, Heng Guo, Guangwei Gao, and Zhanyu Ma. Self-supervised selective-guided  
 123 diffusion model for old-photo face restoration. In *NeurIPs*, 2025.
- 124 [16] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old  
 125 photos back to life. In *CVPR*, pages 2747–2757, 2020.
- 126 [17] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by  
 127 residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.



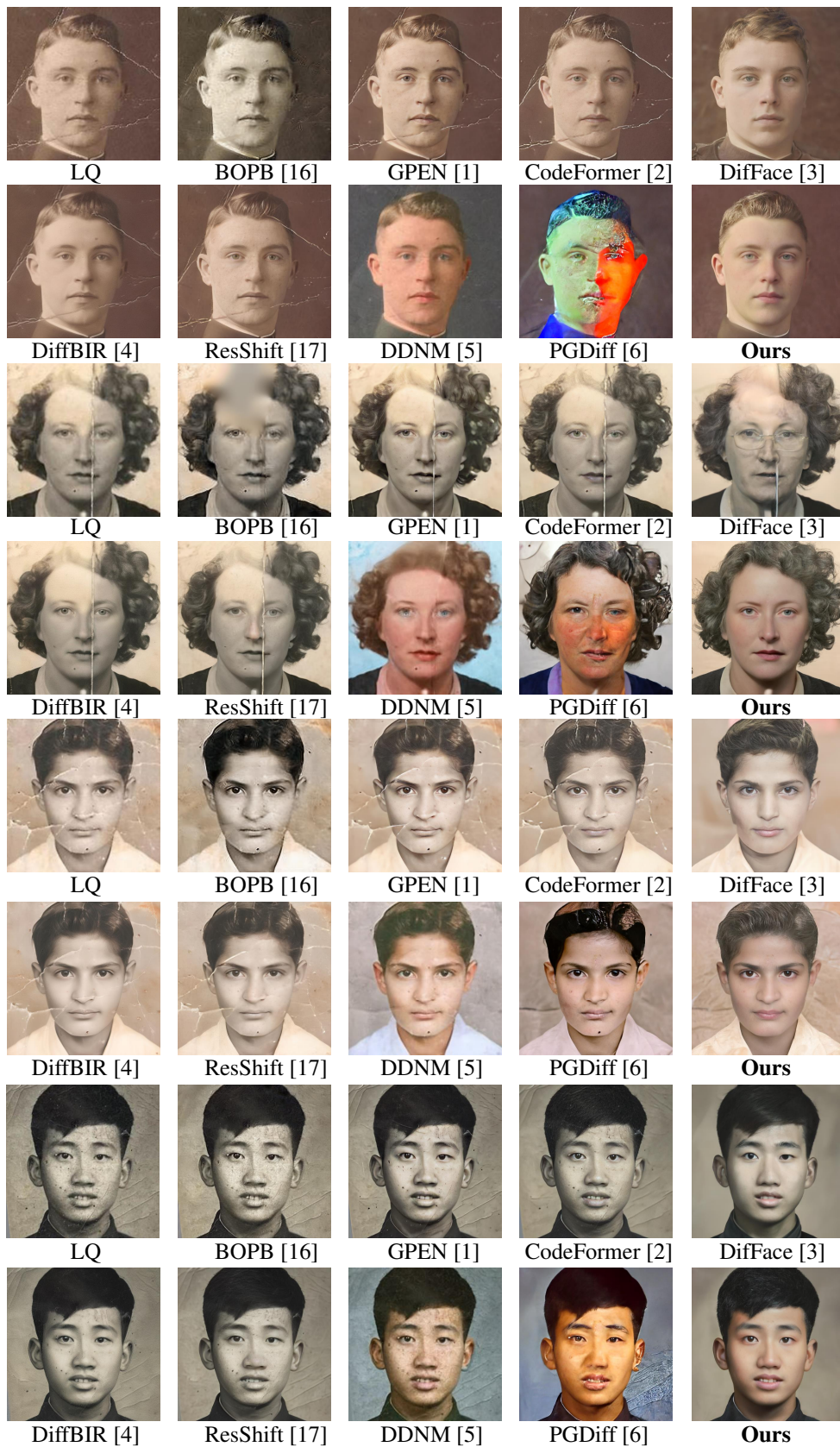


Figure S5: Qualitative comparisons with existing methods on our VintageFace.

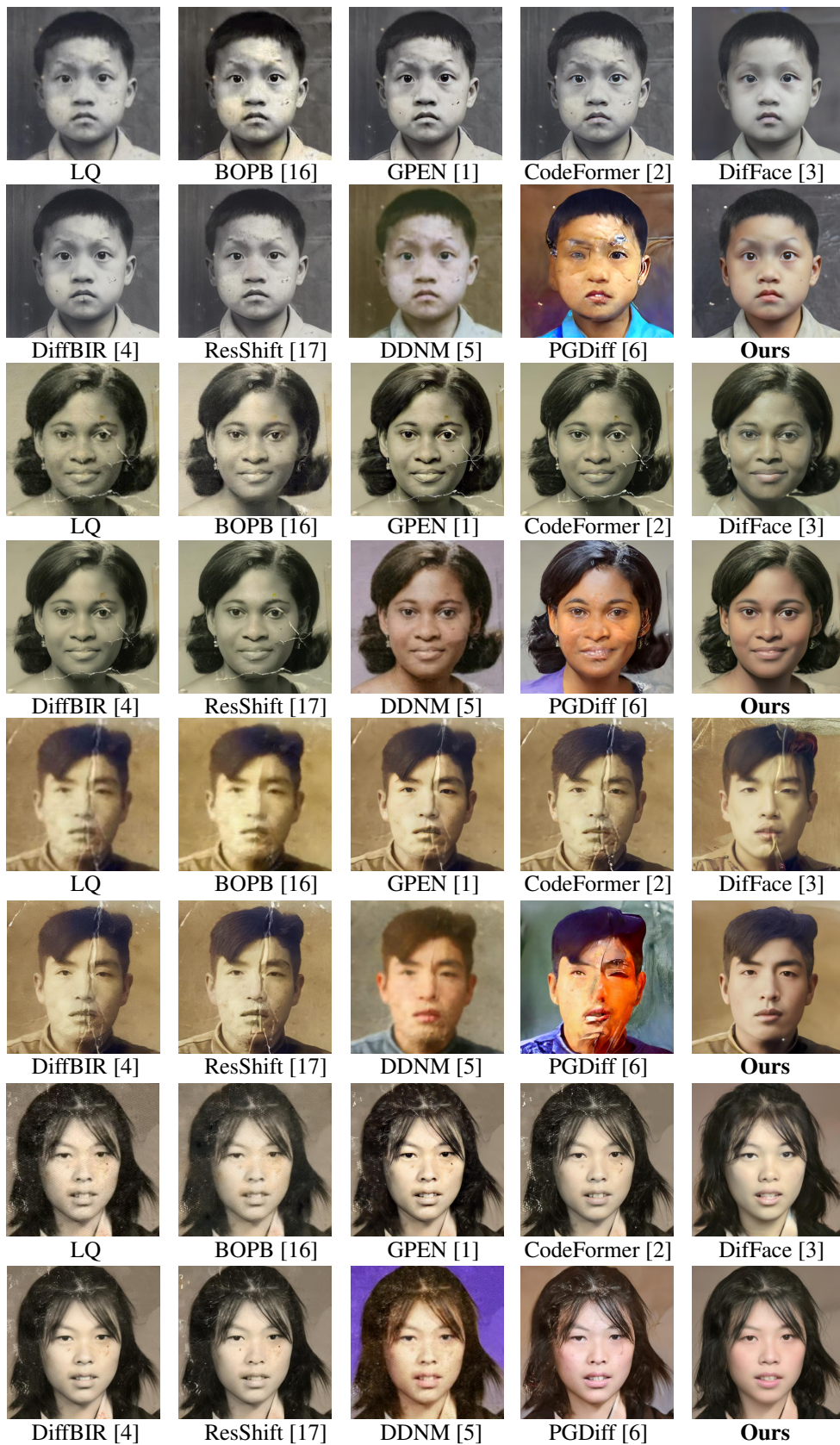


Figure S6: Qualitative comparisons with existing methods on our VintageFace.



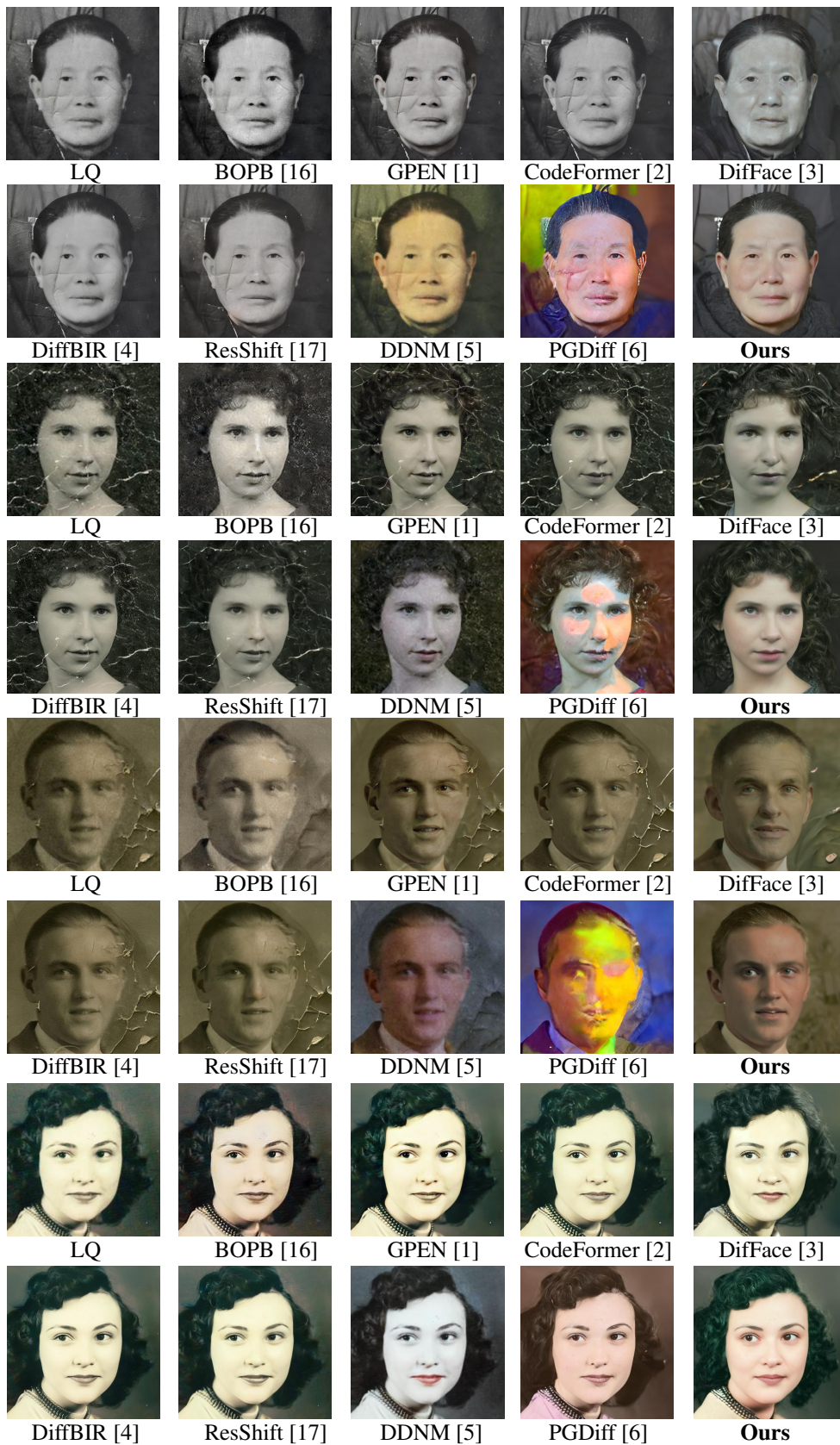


Figure S7: Qualitative comparisons with existing methods on our VintageFace.



Figure S8: Qualitative comparisons with existing methods on WebPhoto-Test and CelebA-Child.