In this document, we present some additional statistics, implementation details, and more examples of the HGFR dataset in Sec. 1 , additional experiments in Sec. 2 , and discuss the broader impact of our work in Sec. 3.

1. Dataset and Implementation Details

This section provides more information about the HGFR dataset. Fig. 1 provide more examples from the HGFR dataset, which are classified by model.

1.1. Dataset Blemish Analysis via CLIP Scoring.

As mentioned in the main paper, Section 3 summarizes the “Face with Blemishes” percentages across several face retouching datasets. These percentages are obtained via CLIP’s cross-modal matching capabilities: for each dataset, we compute the cosine similarity between image embeddings and the text prompts “face with blemish” and “face without blemish.” An image is categorized as blemished if its similarity to “face with blemish” exceeds that to “face without blemish.”

1.2. Prompt Generation

we introduce a systematic method for generating descriptive prompts that integrate multiple facial and demographic attributes. The process begins by defining several categorical variables, each representing a specific attribute of a subject. For example, we maintain lists for age (e.g., teenager, adult, elderly), gender (e.g., male, female), beauty levels (conditioned on gender, such as handsome for males and beautiful for females), hair styles, skin conditions, and ethnicities. In addition, detailed facial features are partitioned into subcategories such as lips, nose, eye size, eye shape, face characteristics, and eyebrows—and a set of expressions is specified.Distribution of prompts shown in Fig. 2

Our implementation leverages Python’s random module to uniformly sample one value from each category, ensuring unbiased and diverse selection. To prevent semantic conflicts, particularly within the facial feature sub-categories, the process restricts each attribute group to a single selection. The selected values are then concatenated to construct a coherent natural language prompt. For instance, the generated output might be:

“A beautiful adult female with bob haircut, few moles on face, East Asian ethnicity, thick lips, tall nose, big eyes, arched eyebrows, and a smile.”

This approach not only guarantees a high degree of variability in the generated prompts but also maintains consistency in the description.

1.3. HGFR Dataset Statistic

Dataset generation was performed on NVIDIA RTX 4090 GPUs, requiring a total of 206 hours, with an average generation time of 45–60 seconds per image pair. The final dataset comprises 25,000 image pairs (7,920 pairs per model, A curated selection of 1,240 high-quality facial images.), capturing extensive variations in facial features, expressions, and blemish types. We partition the dataset into 80% for training, 10% for validation, and 10% for testing, ensuring robust evaluation across diverse scenarios.

1.4. Data Generation Pipeline

We present a systematic three-stage framework to synthesize a diverse facial image dataset with controlled imperfections. First, three pre-trained diffusion models[1][2][3] generate 23,760 base portraits using 7,920 combinatorially designed prompts spanning 7 attributes (age, gender, aesthetics, hairstyle, skin tone, facial morphology). Second, a FaceNet segmentation model [3] isolates 11 facial regions, enabling localized blemish synthesis via 1 Low-Rank Adaptation (LoRA)-tuned model [1] that iteratively inject moles, freckles, and acnes across cheeks/forehead/nose under intensity control (mild/moderate/severe). The process is shown in the Fig. 3

1.5. License and Ethics

As mentioned in the main paper, the HGFR dataset consists of 1,240 high-definition raw portrait photos from the Unsplash website[4], which grants us an irrevocable, nonexclusive, worldwide copyright license[5] to download, copy, modify, distribute, perform, and use photos from Unsplash for free, including for commercial purposes, without permission from or attributing the photographer or Unsplash.

2. Additional Experiments

For fair comparisons among methods with varying inference capabilities, all images are padded and resized to 1024 × 1024 during training and evaluation.

2.1. Extended Ablation Study

To assess the impact of the transformer T components, we conduct ablation studies on its three key modules: frequency-aware dynamic aggregation (FADA) , space domain projection(SDP) and selective frequency feed-forward etwork (SFFN). We consider three variants with MFR: (i) removing the frequency-aware dynamic aggregation (FADA), (ii) removing the space domain projection (SDP), and (iii) disabling frequency domain processing in SFFN. These three settings are denoted R1, R2 and R3, respectively.As reported in Tab. 1 , each component contributes significantly to the removal performance of the blemishes in HGFR + FFHQR.

[1] https://civitai.com/models/25694/epicrealism

[2] https://civitai.com/models/4201?modelVersionId=501240

[3] https://civitai.com/models/133005/juggernaut-xl

[4] https://unsplash.com

[5] https://unsplash.com/license

The soft mask generation branch (SMGB) localizes target regions for precise retouching. As demonstrated in Fig. 4 , without SMGB the model's attention drifts toward non-target areas such as hair and background, leading to degraded performance. In contrast, SMGB-guided processing ensures that retouching is confined to the intended regions, thereby preserving the integrity of non-target areas.

Metric	R1	R2	R3	Ours
PSNR↑	47.58	46.73	48.53	**49.15**
SSIM↑	0.989	0.983	0.991	**0.994**
LPIPS↓	0.0128	0.0153	0.0105	**0.0094**

Table 1. Quantitative ablation experiments on HGFR + FFHQR.

2.2. Parameters Analysis

We conduct a comprehensive analysis on the multi resolution fusion (MRF) module by varying its depth parameter l (number of hierarchical levels). As shown in Table X, increasing l from 1 to 2 yields consistent improvements across all metrics - PSNR increases by 0.33dB (48.82→49.15), SSIM improves from 0.993 to 0.994, and LPIPS decreases by 9.6% (0.0104→0.0094). However, further increasing to l=3 leads to performance saturation with slight degradation in LPIPS (0.0099), suggesting diminishing returns from deeper hierarchies. The performance continues to drop marginally at l=4 (PSNR: 49.03, LPIPS: 0.0102), indicating potential over-parameterization. This reveals a critical trade-off: while deeper MRF structures initially enhance feature aggregation, excessive depth may introduce redundant computation and optimization difficulties. Based on these observations, we empirically set l=2 as the optimal configuration, achieving the best balance between representation capacity and computational efficiency.

Setting	PSNR↑	SSIM↑	LPIPS↓
l = 1	48.82	0.993	0.0104
l = 2	**49.15**	**0.994**	**0.0094**
l = 3	49.08	0.994	0.0099
l = 4	49.03	0.994	0.0102

Table 2. Comparison of evaluation metrics of our model in the case of different resolutions on FFHQR+HGFR.

2.3. More Visualization Results

More comparison examples on the FR-wild datasets are shown in Fig. 5 . It can be seen that our method performs favorably against the others.

3. Broader Impact

Our synthetic dataset generation pipeline addresses privacy concerns in facial data collection. HGFR will be publicly released to support future research in face retouching and synthetic data generation. Prompts will be publicly released to support research in face generation.

References

[1] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. 1

[2] Biwen Lei, Xiefan Guo, Hongyu Yang, Miaomiao Cui, Xuansong Xie, and Di Huang. Abpn: adaptive blend pyramid network for real-time local retouching of ultra high-resolution photo. In *CVPR*, pages 2108–2117, 2022. 6

[3] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Facex: Understanding face attribute classifiers through summary model explanations. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 758–766, 2024. 1

[4] Alireza Shafaei, James J Little, and Mark Schmidt. Autoretouch: Automatic professional face retouching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 990–998, 2021. 6

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018. 6

[6] Xue Wen, Lianxin Xie, Le Jiang, Tianyi Chen, Si Wu, Cheng Liu, and Hau-San Wong. Retouchformer: semi-supervised high-quality face retouching transformer with prior-based selective self-attention. In *AAAI*, pages 5903–5911, 2024. 6

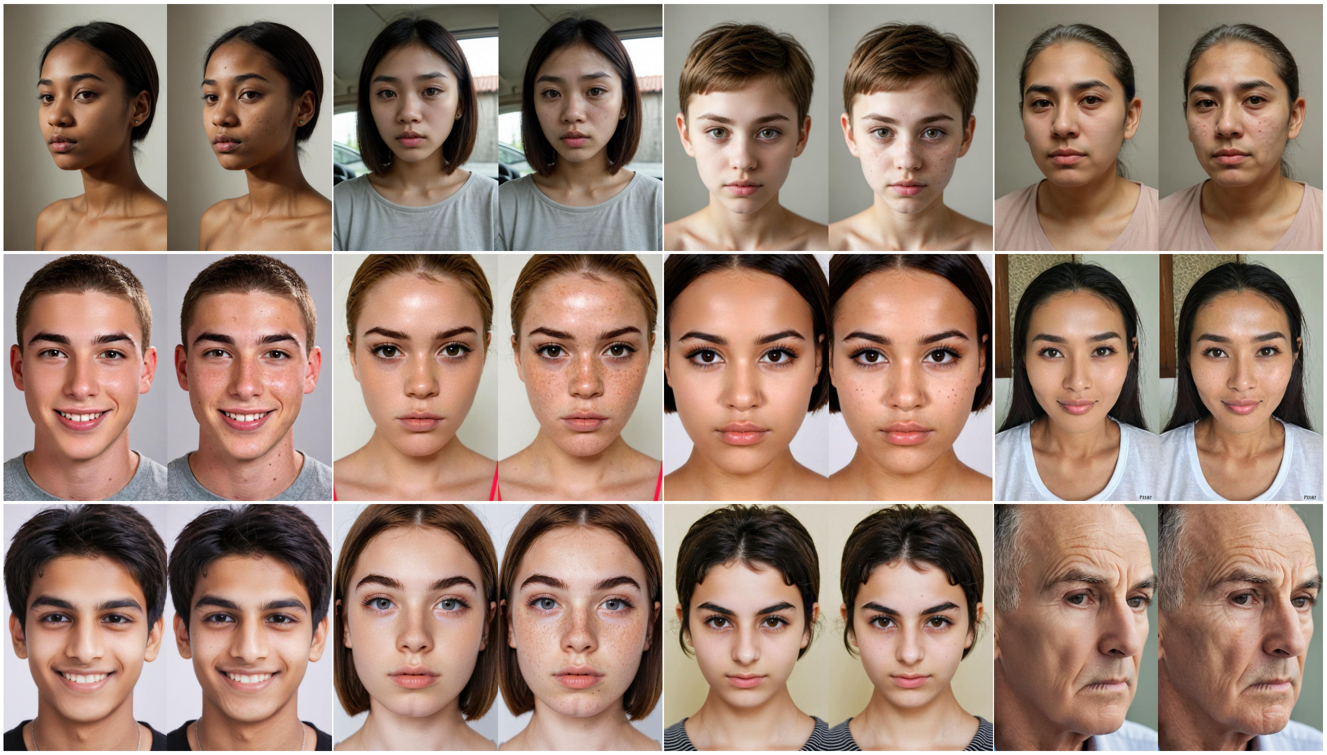


Figure 1. Examples classified by Model of generated face from the HGFR Dataset (zoom in for a better view). From top to bottom: epicrealism,juggenaut,realisticvision.From left to right:face without blemishes, face with blemished.

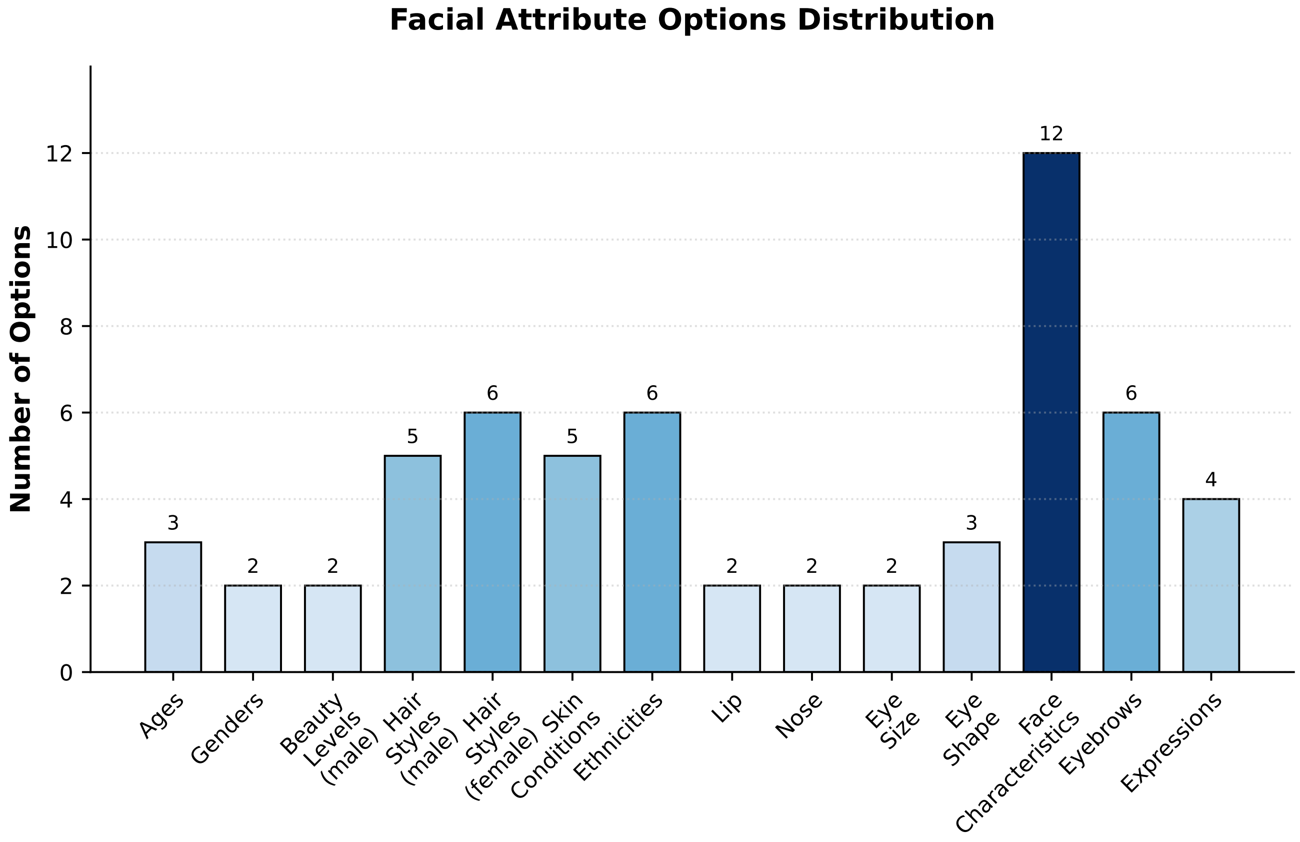


Figure 2. Distribution of prompts.

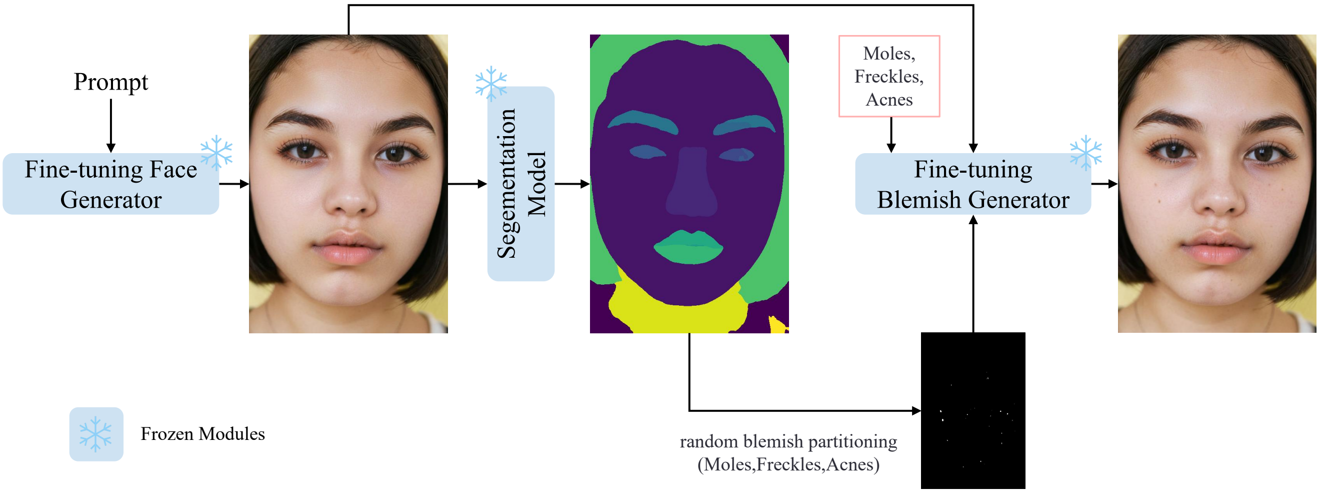

Prompt
Fine-tuning Face Generator
Segementation Model
Moles, Freckles, Acnes
Fine-tuning Blemish Generator
random blemish partitioning (Moles,Freckles,Acnes)
Frozen Modules

Figure 3. Illustration of the data generation process.

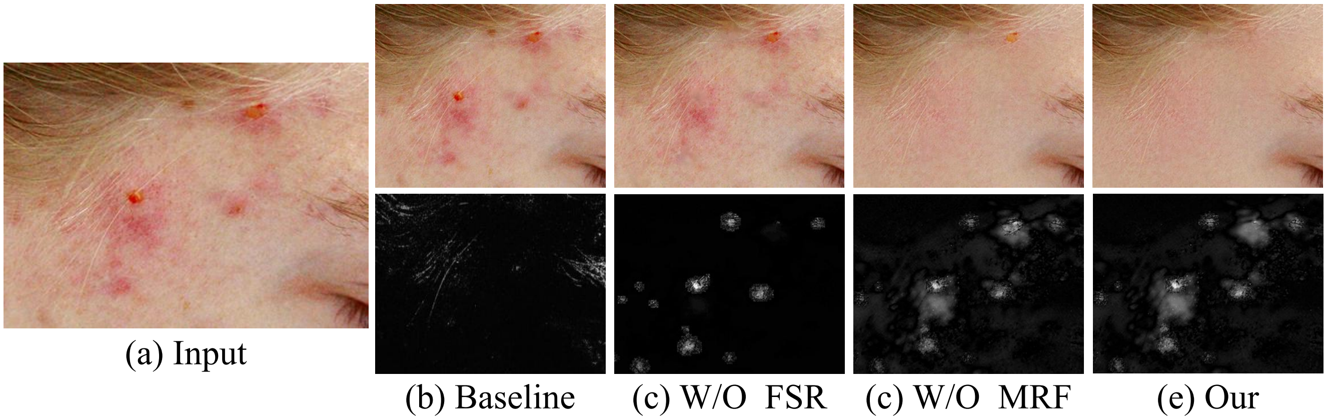
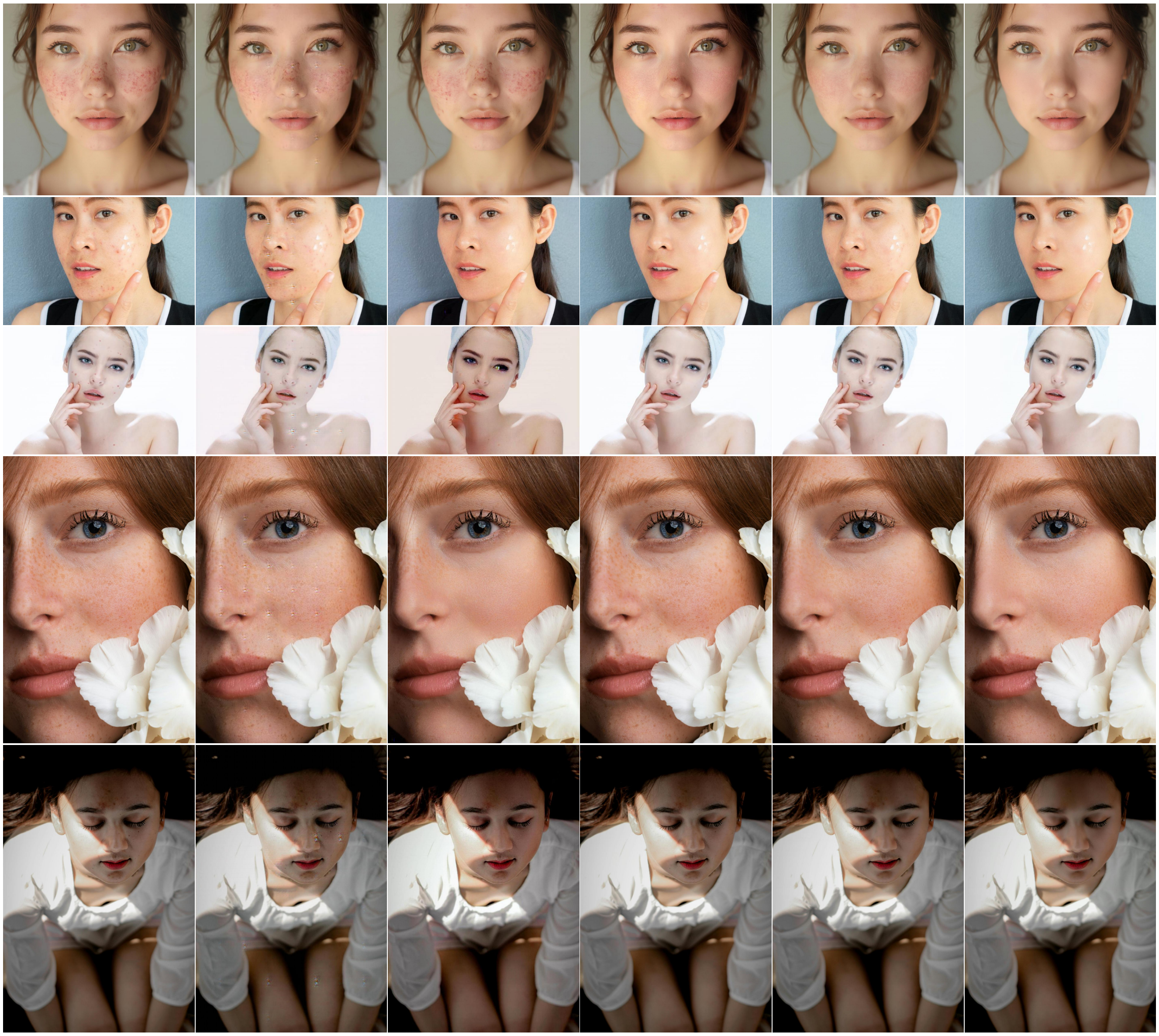


Figure 4. Ablation study toward SMGB, FSR and MRF on HGRF+FFHQR. (a) Input image; (b) Result on baseline; (c) Result on baseline+SMGB; (d) Result on baseline+SMGB+FSR; (e) Result on baseline+SMGB+FSR+MRF. The masks presented in the bottom of the last four columns show the changing area relative to the input.



(a) Input (b) AutoRetouch (c) pix2pixHD (d) ABPN (e) RetouchFormer (f) Ours

Figure 5. Qualitative comparisonon FFHQR and HGRF(zoomin for abetter view): (a) original images, (b) Autoretouch [4] , (c) pix2pixHD [5] (d) ABPN [2] , (e) RetouchFormer [6] , (f) Ours.