

NTIRE 2025 Challenge on Real-World Face Restoration: Methods and Results

Supplementary Material

Zheng Chen*	Jingkai Wang*	Kai Liu*	Jue Gong*	Lei Sun*	Zongwei Wu*
Radu Timofte*	Yulun Zhang*†	Jianxing Zhang	Jinlong Wu	Jun Wang	Zheng Xie
Hakjae Jeon	Suejin Han	Hyung-Ju Chun	Hyunhee Park	Zhicun Yin	
Junjie Chen	Ming Liu	Xiaoming Li	Chao Zhou	Wangmeng Zuo	Weixia Zhang
Dingquan Li	Kede Ma	Yun Zhang	Zhuofan Zheng	Yuyue Liu	Shizhen Tang
Zihao Zhang	Yi Ning	Hao Jiang	Wenjie An	Kangmeng Yu	Chenyang Wang
Kui Jiang	Xianming Liu	Junjun Jiang	Yingfu Zhang	Gang He	Siqi Wang
Kepeng Xu	Zhenyang Liu	Changxin Zhou	Shanlan Shen	Yubo Duan	
Yiang Chen	Jin Guo	Mengru Yang	Jen-Wei Lee	Chia-Ming Lee	
	Chih-Chung Hsu	Hu Peng	Chunming He		

A. More Challenge Methods and Teams

A.1. UpHorse

Description. The proposed diffusion-prior-based implicit representation joint face restoration method, DSS (Fig. 1), utilizes three SOTA restoration and SR models for the joint restoration of degraded images. Specifically, the process begins with fine-tuning DiffBIR [7] to recover the fundamental structure of the image. Next, StableSR [10] is employed to precisely align the facial regions in real-world scenarios. Finally, SUPIR [12] is used to further enhance the facial details. By effectively leveraging the strengths of each model, DSS achieves comprehensive, high-quality restoration of degraded images, ensuring superior performance across various aspects of the restoration process.

Implementation Details. For images with varying degradation levels across different scenes, they apply three separate models for individual restoration. Experimental Fig. 2 results show that, in the initial restoration phase, DiffBIR [7] achieves superior results, while StableSR [10] and SUPIR [12] tend to restore more details. Based on these observations, they propose a three-stage face restoration framework: first, DiffBIR [7] is used for preliminary restoration to address the basic structure of the image; then, StableSR [10] is employed to further align the facial fea-

tures with real human faces; finally, SUPIR [12] is applied to enhance facial details, thereby achieving comprehensive and high-quality restoration.

To improve the degradation removal performance of DiffBIR [7] in the first stage, they fine-tune the first-stage SwinIR [6] of DiffBIR [7] using 50,000 FFHQ [3] images. Specifically, they crop the input images to 512×512 , set the learning rate to $1e-4$, and train for 150K iterations on an NVIDIA 4090 GPU. Considering the varying degradation levels in the test data, they divided the degradation settings into different ranges to better accommodate different levels of image degradation. The loss function used is the Mean Squared Error (MSE) loss Equation. 1.

$$I_{SW} = SW(I_{lq}), L = \|I_{SW} - I_{hq}\|_2^2, \quad (1)$$

Where SW refers to SwinIR, I_{SW} represents the output of SwinIR, I_{lq} denotes the low-quality image, and I_{hq} denotes the high-quality image.

In terms of the cfg scale parameter setting, larger values result in lower fidelity and more details. Given the input data distribution requirements of subsequent models, they set the cfg scale to 6 during the restoration process with DiffBIR [7], thus achieving a balance between the restoration quality and the input requirements for subsequent detail enhancement. Other parameters remain unchanged from the original DiffBIR [7]. Experimental Fig. 3 results demonstrate that the fine-tuned DiffBIR [7] achieves significantly improved performance in degradation removal.

The experiments in Fig. 4 and Fig. 5 (“+” and “++” indicate cumulative improvements over the previous method) show that in the second stage, using StableSR [10] on top of the first-stage restoration achieves further alignment with

*Zheng Chen, Jingkai Wang, Kai Liu, Jue Gong, Lei Sun, Zongwei Wu, Radu Timofte, and Yulun Zhang are the challenge organizers, while the other authors participated in the challenge. Section B in the supplementary materials contains the authors’ teams and affiliations. NTIRE 2025 webpage: <https://cvlai.net/ntire/2025>. Code: https://github.com/zhengchen1999/NTIRE2025_RealWorld_Face_Restoration.

†Corresponding author: Yulun Zhang. yulun100@gmail.com



Figure 1. **Team UpHorse**. Overall pipeline.

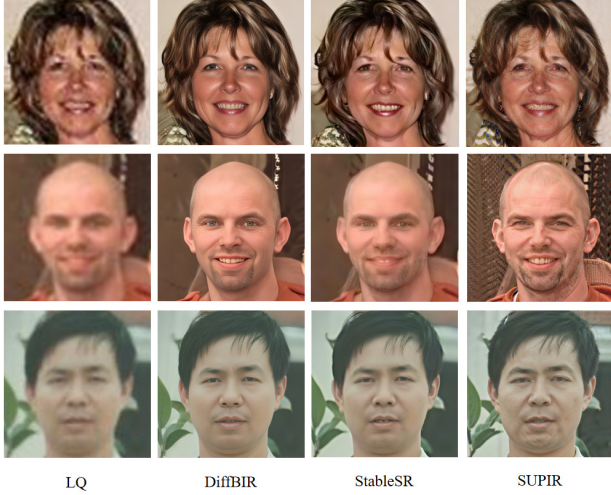


Figure 2. **Team UpHorse**. The results in initial restoration stage.

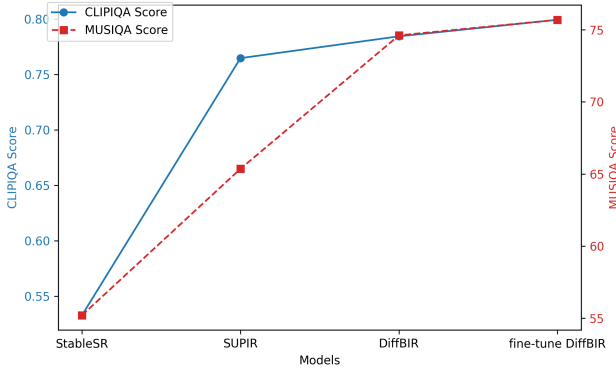


Figure 3. **Team UpHorse**. The scores in initial restoration stage.

real-world faces and enhances fine details. They kept the inference settings of StableSR [10] unchanged. In the third stage, using SUPIR [12] based on the second stage further enhances the texture details of the face.

Given SUPIR’s [12] powerful super-resolution capabilities and considering limited computational resources, they removed the language-guided restoration module LLaVA [8]. Instead, they upsampled the input images by a factor of 3 and then downsampled them to create relatively low-quality input images, which are specifically targeted by SUPIR [12]. This adjustment led to an improvement in the



Figure 4. **Team UpHorse**. The restoration results at each stage.

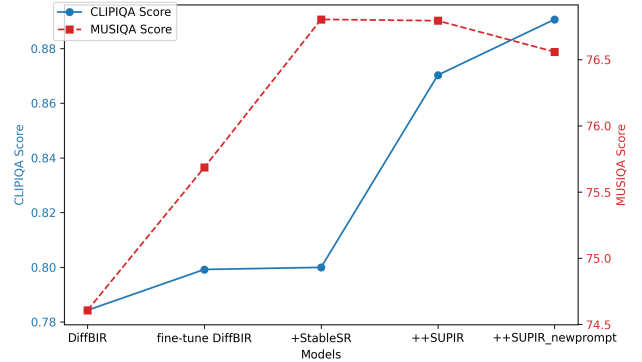


Figure 5. **Team UpHorse**. The scores at each restoration stage.

results after using SUPIR [12] for restoration. Additionally, they experimented with various forward and reverse prompt combinations to replace the original prompt, leading to further breakthroughs in the results.

A.2. CX

Description. Their network design, shown in Fig. 6, is exactly the same as DiffBIR. They only performed data filtering and processing on FFHQ and then fine-tuned based on the pre-trained model of DiffBIR.

Implementation Details. The proposed method achieves a

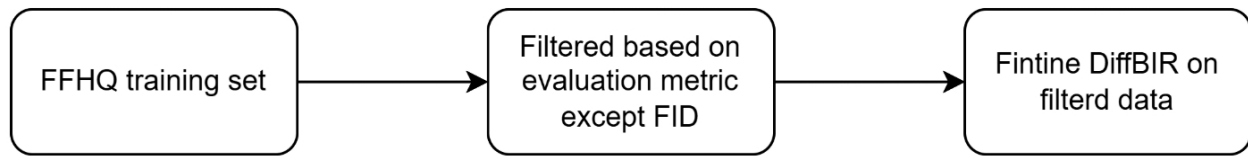


Figure 6. Team CX.

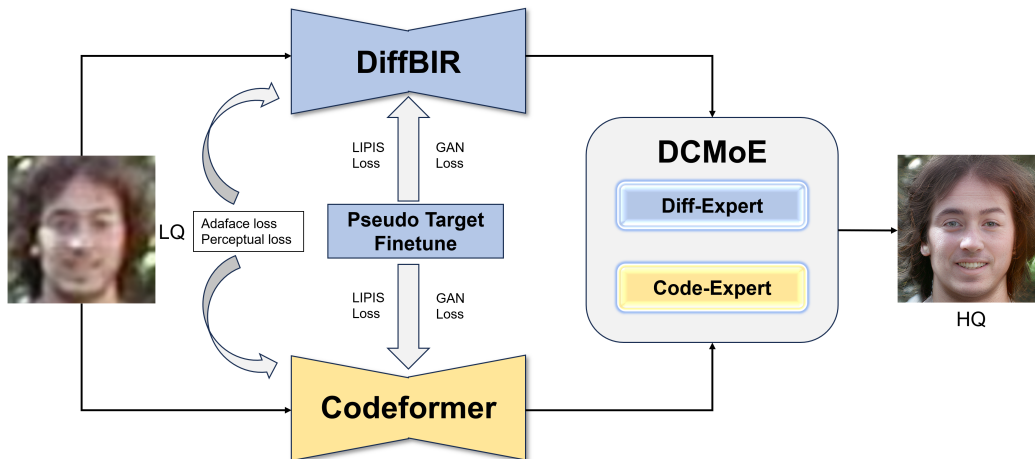


Figure 7. Team AIIALab.

runtime of 5 seconds per image (512x512) on an NVIDIA RTX 4090. It utilizes the DiffBIR model (version: v1_face) and is trained exclusively on the FFHQ dataset. The training process uses Adam as the optimizer with a learning rate of $1.17e-6$ for 80k iterations, conducted on a 512x512 resolution using 5 NVIDIA A5000 GPUs with a batch size of 30 (6 per GPU). During testing, 50 steps are performed with an empty positive prompt and a negative prompt of “low quality, blurry, low-resolution, noisy, unsharp, weird textures”, without using any image captioner for guidance. The method shows a 0.06 improvement in the final score: the weighted score of original v1_face DiffBIR is 3.97, after finetuning on filtered FFHQ, the weighted score is 4.03. The solution is based on filtering the FFHQ training set.

A.3. AIIALab

Description. They address the challenge of restoring high-quality facial images from low-quality real-world inputs while preserving ID consistency. To achieve high-fidelity reconstruction with enhanced visual quality, they propose a dual-model adaptive Mixture-of-Experts (MoE) architecture that synergistically integrates the high-quality generative priors of diffusion models with the identity-preserving capabilities of transformer networks. Specifically, their

framework combines the complementary strengths of DiffBIR (for quality enhancement through diffusion processes) and CodeFormer (for structural fidelity via transformer-based facial priors), implementing dynamic feature fusion through their novel adaptive gating mechanism.

Implementation Details. DiffBIR[7] proposes a two-stage framework for blind image restoration that systematically decouples degradation removal and content reconstruction. In the first stage, they employ dedicated restoration modules to eliminate image-independent degradations (e.g., noise, blur), leveraging existing or custom-trained models (e.g., swinir) tailored for specific distortion types. The second stage focuses exclusively on semantic-aware content regeneration through their designed generation module, which operates solely on the purified image content from the first stage, thereby avoiding interference from residual artifacts. Crucially, they implement independent optimization strategies for both stages while maintaining inter-stage compatibility. To achieve dynamic quality control, they further develop a training-free adaptive guidance mechanism that spatially modulates restoration intensity during the diffusion sampling process, enabling a precise balance between perceptual quality and structural fidelity across different image regions. This architecture provides a unified yet flexible

solution that combines task adaptability with stable performance across diverse blind restoration scenarios.

Codeformer[13] begins by integrating vector quantization principles to establish a semantic-aware codebook through a self-reconstruction-driven pre-training process, where they first train a quantized autoencoder to learn discrete latent representations and their corresponding decoder. Building upon this learned codebook prior, they then design a Transformer-based architecture to precisely predict optimal code combinations directly from degraded facial inputs, enabling targeted restoration of missing facial details. To dynamically balance perceptual quality and identity preservation, they further develop a controllable feature transformation module that adaptively adjusts feature representations during the restoration process. The entire system follows a three-stage progressive training strategy: codebook construction, code prediction optimization, and fidelity-quality adaptive refinement, ensuring systematic alignment between prior knowledge extraction and task-specific restoration objectives.

Their method’s pipeline is shown in Fig. 7. Their framework introduces an adaptive Mixture-of-Experts (MoE) module to dynamically integrate the outputs of DiffBIR and Codeformer, optimizing the selection of high-fidelity and visually plausible restored facial images. Specifically, the MoE module operates through three key mechanisms. **1)** The Diffusion Expert leverages iterative denoising to recover fine-grained details but may introduce identity shifts under severe degradations, while the Codebook Expert utilizes a pretrained vector-quantized autoencoder to enforce identity consistency through discrete code prediction, albeit with limited detail recovery capability. **2)** A lightweight gating network analyzes multi-scale degradation features (e.g., noise distribution, blur kernels) and semantic cues (facial landmarks, identity embeddings) from the low-quality input. It predicts dynamic weights w_{diff} and $w_{\text{code}} = 1 - w_{\text{diff}}$ via a Sigmoid-activated MLP, prioritizing DiffBIR for noise/blur-dominated inputs and Codeformer for identity-critical scenarios (e.g., extreme low-resolution or occlusions). **3)** The final output is computed as $I_{\text{final}} = w_{\text{diff}} \cdot I_{\text{diff}} + w_{\text{code}} \cdot I_{\text{code}}$, jointly optimized by a hybrid loss combining pixel-level reconstruction (\mathcal{L}_{L1}), identity preservation (\mathcal{L}_{ID}), and adversarial training (\mathcal{L}_{adv}).

Training datasets. They use 70k FFHQ as their train datasets, no other data. All images are randomly cropped to 512×512 during training. And they used data augmentation methods such as random rotation and flipping to expand the diversity of the dataset.

Training strategy. In the DiffBIR [7], they train the restoration module for 150k iterations (batch size=96). Then, they adopt Stable Diffusion 2.1-base1 as the generative prior, and finetune the proposed IRControlNet for 80k iterations (batch size=256). Adam is used as the optimizer. The learn-

ing rate is set to 10^{-4} for the first 30k iterations and then decreased to 10^{-5} for the following 50k iterations.

In the codeformer [13], they represent a face image of 512×512 as a 16×16 code sequence. For all stages of training, they use the Adam optimizer with a batch size of 16. They set the learning rate to 8×10^{-5} for stages I and II, and adopt a smaller learning rate of 2×10^{-5} for stage III. The three stages are trained with 1.5M, 200K, and 20K iterations, respectively.

In the training process of the above two models, they also added perceptual loss and AdaFace loss to obtain high-fidelity and high-quality facial images, with weights of 0.01 for both losses.

After that, they also apply the DT-BFR[5] method to finetune the SwinIR model and Codeformer model (SwinIR model is used as the DiffBIR stage 1 model). They first generate pseudo targets using a diffusion model and then use the generated targets to fine-tune the two pre-trained restoration models. For fine-tuning the SwinIR model, they set the weights of the losses to be $\lambda_{\text{LPIPS}} = 0.1$ and $\lambda_{\text{GAN}} = 0.1$ for all the experiments. For CodeFormer, they follow their training setup and empirically found that only adopting their code-level losses to optimize the code prediction module and the VQ-GAN encoder gives better fine-tuning performance than the image-level losses.

A.4. ACVLab

Description. As Fig. 8, they proposed a two-stage restoration approach for blind face restoration that combined GFPGAN [11] and DiffBIR [7] to enhance real-world degraded facial images. GFPGAN is applied to the cropped face images to perform coarse restoration. This step effectively reconstructs missing facial details and provides an initial enhancement of the global structure while maintaining identity consistency. Afterwards, the output of GFPGAN is fed into DiffBIR, a diffusion-based blind image restoration model for fine-grained-level restoration and refinement.

Implementation Details. The training process utilizes the FFHQ [3] and FFHQ-R [9] datasets at a resolution of 512×512 pixels. They performed image degradation realistically via the standard degradation pipeline offered by the organizer on CodaLab. The degraded input images undergo initial alignment and normalization to ensure consistency across different samples. RetinaFace [1] is applied to locate facial regions, which are then cropped and resized to a standard resolution.

For training GFPGAN, their training utilizes several key optimization parameters. Adam optimizer [4] is employed for both generator and discriminator networks with a learning rate of 2×10^{-3} . A MultiStepLR scheduler is implemented with milestones at 600,000 and 700,000 iterations (gamma 0.5). The total training process consists of 800,000 iterations. The discriminator is trained at a frequency of

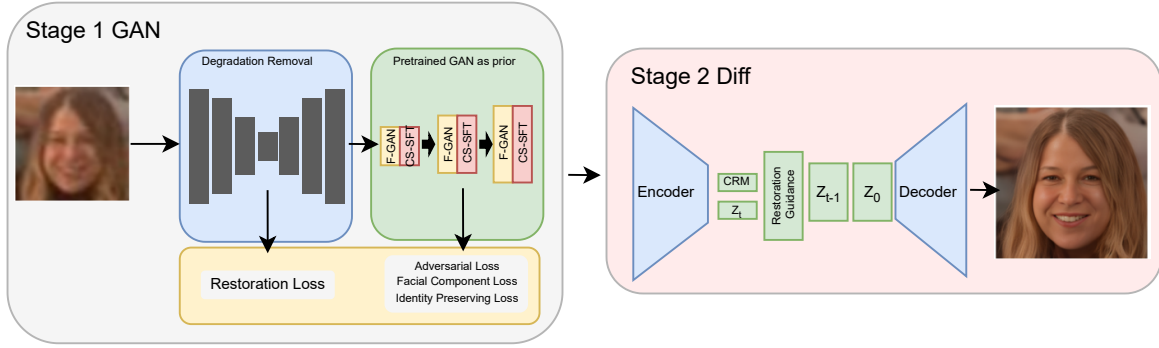


Figure 8. Team ACVLab.

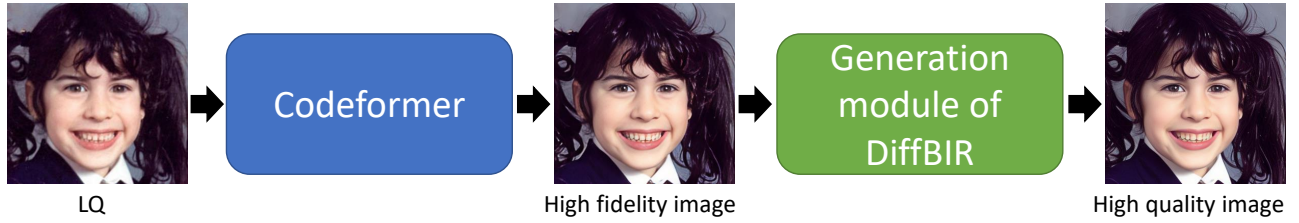


Figure 9. Team HH.

once per generator update. As for the objective function, they follow standard setting from the official repository of GFPGAN.

A.5. HH

Description. Their method (Fig. 9) relies on DiffBIR [7] and CodeFormer [13]. DiffBIR is a two-stage image restoration model. In the first stage, a base restoration model is responsible for generating a high-fidelity preliminary result. In the second stage, a diffusion model based on ControlNet conditional control is employed to add high-frequency details to the output of the first stage. they utilize CodeFormer as the first-stage model of DiffBIR, combining it with DiffBIR’s second-stage model to achieve high-fidelity and high-quality portrait restoration.

Implementation Details. Their method integrates *CodeFormer* and *DiffBIR*, two existing approaches, and thus does not require retraining. The training strategies for these methods can be referenced in their respective papers.

A.6. Fustar-fsr

Description. They present their model named Incremental Face Restoration Model. This model has a hierarchical face restoration framework, which integrates progressive generative diffusion models and face prior guidance, and is tailored to address the complex degradation issues in

real-world scenarios. By leveraging the unique capabilities of PGDiff and GFPGAN, and incorporating post-stage image enhancement, they aim to achieve superior restoration results.

Implementation Details. Many current blind face restoration methods are quite excellent, which is beneficial for us to refer to them and achieve real-world face super-resolution. For instance, methods based on Generative Adversarial Networks (GANs), like GFPGAN (Generative Facial Prior-GAN), leverage the rich and diverse prior knowledge contained in pre-trained face GANs (such as StyleGAN2). This prior knowledge encompasses various features and structural information of the face, including the distribution of facial features, texture patterns, and colors, to guide the face restoration process. In the context of real-world face super-resolution, they can draw on this prior knowledge to assist the model in better understanding the structure and features of the face, thereby generating more reasonable and natural details when upscaling the image. For example, DiffBIR: It decomposes the blind image restoration problem into two sub-problems: degradation removal and detail generation. A two-stage framework is proposed. In the first stage, the restoration module trained by the MSE loss function is used to remove most of the degradation contained in the image, obtaining a clean but smooth image lacking local texture details. In the second stage, the

ControlNet module is used to leverage the generative power of Stable Diffusion to compensate for lost texture details or semantic information. Meanwhile, a controllable module without the need for additional training-latent image guidance, is introduced to balance image quality and fidelity. And CodeFormer: It transforms the blind face restoration problem into a code prediction task. A discrete codebook is used to represent the local features of high-quality face images, and low-quality face images are mapped to the code space of high-quality images. By introducing the codebook lookup Transformer (CLT), it combines the advantages of discrete codebooks and Transformers. The self-attention mechanism is adopted to model the global and local features in images. And a hybrid training strategy is used to combine reconstruction error and perceptual loss to ensure that the model can not only remove noise and blur but also retain the realism and naturalness of the image during the restoration process.

After referring to those models, they primarily propose a hierarchical face restoration framework based on a progressive generative diffusion model and face prior guidance, and conduct further image enhancement in the subsequent stage. As the first-step processing method, they present a multi-stage cascaded framework. Specifically, this method successfully integrates the temporally controllable generation ability of the progressive generative diffusion model (PGDiff) with the structural semantic prior of the face prior generation model (GFPGAN), achieving progressive face restoration from coarse-grained to fine-grained levels.

(1) *Utilization of PGDiff’s multi-scale diffusion guidance.* They make use of PGDiff’s conditional-guided diffusion mechanism. During the diffusion process, they introduce time-step-adaptive semantic constraints. By dynamically fusing low-quality inputs (LQ) with pre-generated face prior features of GFPGAN at different diffusion stages (ranging from the high-noise level where $t = T$ to the low-noise level where $t = 0$), they realize progressive restoration from the global structure to local details.

(2) *Semantic prior embedding of GFPGAN* They regard GFPGAN as a semantic prior generator. Through adaptive instance normalization (AdaIN), it encodes its pre-trained knowledge of face structure (such as the distribution of facial features and texture patterns) into conditional vectors. These vectors are then input as class-conditional (class-cond) information during the reverse sampling process of PGDiff, constraining the diffusion model to generate high-resolution images that conform to the statistical characteristics of real faces.

(3) *Subsequent-stage image enhancement processing* To obtain more vivid images in terms of visual perception, they draw on the methods in the HGGT paper and design an image enhancement model aiming to further enhance the perceptual quality of the restored face images. Specifically,

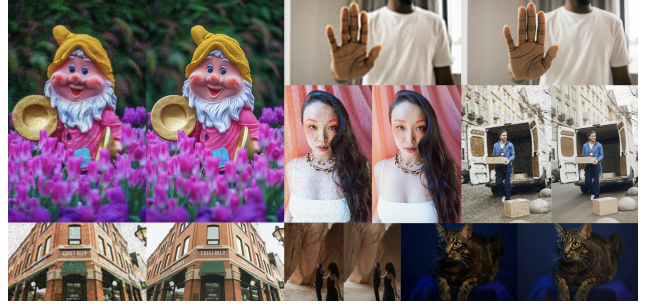


Figure 10. Team IPCV

they train an SR model named ELAN with L1 loss, perceptual loss, and adversarial loss, and apply it to HQ images to further improve the perceptual quality of these images. They use the commonly used DF2K as the training set and degrade the HR images. Before inputting the degraded images into the model training, they resize them to the size of HR images. Additionally, they remove the upsampling layer in the model because, in order to obtain face images with enhanced perceptual quality, the input and output need to be the same size.

Training dataset. FFHQ is a high-quality image dataset of human faces, originally created as a benchmark for generative adversarial networks (GAN). The dataset consists of 70,000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity, and image background. It also has good coverage of accessories such as eyeglasses, sunglasses, hats, etc. The images were crawled from Flickr, thus inheriting all the biases of that website, and automatically aligned and cropped using dlib. Only images under permissive licenses were collected. Various automatic filters were used to prune the set, and finally, Amazon Mechanical Turk was used to remove the occasional statues, paintings, or photos of photos. The participants are allowed to use extra data for training.

Training strategy. They use the Adam optimizer for training their models. The optimizer’s parameters are set as follows: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$. These values have been empirically found to provide stable convergence during training.

A.7. Night Watch

Description. They use the initial weights provided by <https://github.com/XPixelGroup/DiffBIR>. Specifically, the original SwinIR was trained on ImageNet-1k with CodeFormer degradation, and IRControlNet was trained on filtered laion2b-en. For detailed training hyperparameters, please refer to the corresponding paper. They fine-tune SwinIR using the FFHQ dataset.

A.8. IPCV

Description. The Flux.1-dev-Controlnet-Upscaler is a deep learning model designed for image upscaling, integrating ControlNet with the Flux.1-dev architecture. It is based on a 12-billion-parameter rectified flow transformer, originally designed for high-quality image generation. The incorporation of ControlNet allows for guided upscaling by conditioning on additional inputs like depth maps, ensuring better structure preservation. The training process involved synthetic data degradation techniques, including Gaussian and Poisson noise addition, image blurring, and JPEG compression, making it robust for real-world image restoration. The model can be used with the diffusers library to enhance low-resolution images, and the implementation details are available on its Hugging Face page [2], shown in Fig. 10. He adjusted the 4x scaling to 1x to maintain the original dimensions while enhancing pixel quality.

Implementation Details. No training, only used inference.

B. Teams and Affiliations

NTIRE 2025 team

Title: NTIRE 2025 Real-world Face Restoration Challenge

Members:

Zheng Chen¹ (zhengchen.cse@gmail.com),
Jingkai Wang¹ (jingkaiwang100@gmail.com),
Kai Liu¹ (normal.kliu@gmail.com),
Jue Gong¹ (g1017325431@gmail.com),
Lei Sun² (leosun0331@gmail.com),
Zongwei Wu³ (zongwei.wu@uni-wuerzburg.de),
Radu Timofte³ (radu.timofte@uni-wuerzburg.de),
Yulun Zhang¹ (yulun100@gmail.com)

Affiliations:

¹ Shanghai Jiao Tong University, China

² INSAIT, Bulgaria

³ University of Würzburg, Germany

AllForFace

Title: Using Divide-and-Conquer for Blind Face Restoration

Members:

Jianxing Zhang¹ (jx2018.zhang@samsung.com), Jinlong Wu¹, Jun Wang¹, Zheng Xie¹, Hakjae Jeon², Suejin Han², Hyung-Ju Chun², Hyunhee Park²

Affiliations:

¹ Samsung R&D Institute China - Beijing (SRC-B)

² Samsung MX(Mobile eXperience) Business

III

Title: Blind Face Restoration with One-Step Diffusion Framework

Members:

Zhicun Yin¹ (cszcyin@outlook.com), Junjie Chen¹, Ming Liu¹, Xiaoming Li¹, Chao Zhou², Wangmeng Zuo¹

Affiliations:

¹ Harbin Institute of Technology

² Shanghai Transsion Co, Ltd

PISA-MAP

Title: PiSA-MAP

Members:

Weixia Zhang¹ (zwx8981@sjtu.edu.cn), Dingquan Li², Kede Ma³

Affiliations:

¹ Shanghai Jiao Tong University

² Pengcheng Laboratory

³ City University of Hong Kong

MiPortrait

Title: MPSR

Members:

Yun Zhang¹ (zhangyun9@xiaomi.com), Zhuofan Zheng¹, Yuyue Liu¹, Shizhen Tang¹, Zihao Zhang¹, Yi Ning¹, Hao Jiang¹

Affiliations:

¹ Xiaomi Inc.

AIIA

Title: A Face Image Restoration Method Applying Pre-trained Models and Test-time Adaptation

Members:

Wenjie An¹ (anwenjie1213@163.com), Kangmeng Yu¹

Affiliations:

¹ Harbin University of Technology

UpHorse

Title: DSS: Implicit Representation-Based Face Restoration with Diffusion Prior

Members:

Yingfu Zhang¹ (zmund0717@gmail.com), Gang He¹, Siqi Wang¹, Kepeng Xu¹, Zhenyang Liu¹

Affiliations:

¹ Xidian University

CX

Title: CX

Members:

Changxin Zhou¹(changxin.zhou@bst.ai), Shanlan Shen¹, Yubo Duan¹

Affiliations:

¹Black Sesame Technologies (Singapore) Pte Ltd

AIIALab

Title: DCMoE-RWFR

Members:

Yiang Chen¹(xantares606@gmail.com), Kui Jiang¹, Jin Guo¹, Mengru Yang¹, Junjun Jiang¹

Affiliations:

¹Harbin Institute of Technology

ACVLab

Title: ACVLab

Members:

Jen-Wei Lee¹(jemmy112322@gmail.com), Chia-Ming Lee¹, Chih-Chung Hsu^{1,2}

Affiliations:

¹Institute of Data Science, National Cheng Kung University

²Institute of Intelligent Systems, National Yang Ming Chiao Tung University

HH

Title: HH

Members:

Hu Peng¹(hup22@mails.tsinghua.edu.cn), Chunming He²

Affiliations:

¹Shenzhen International Graduate School, Tsinghua University

²Duke University

Fustar-fsr

Title: Incremental Learning-Face Restoration Model

Members:

Tingyi Mei¹(18084795694@163.com), Qizhao Lin¹, Jialiang Chen¹

Affiliations:

¹Fujian Normal University

Night Watch

Title: Night Watch

Members:

Kepeng Xu¹(kepengxu11@gmail.com), Siqi Wang¹, Yingfu Zhang¹, Zhenyang Liu¹, Gang He¹

Affiliations:

¹Xidian University

IPCV

Title: Fluxoration

Members:

Jameer Babu Pinjari¹(jameer.jb@gmail.com)

Affiliations:

¹Independent Researcher

References

- [1] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 4
- [2] JasperAI. Flux.1-dev-controlnet-upscaler. <https://huggingface.co/jasperai/Flux.1-dev-Controlnet-Upscaler>, 2025. 7
- [3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 4
- [4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4
- [5] Tianshu Kuai, Sina Honari, Igor Gilitschenski, and Alex Levinstein. Towards unsupervised blind face restoration using diffusion prior. *arXiv preprint arXiv:2410.04618*, 2024. 4
- [6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 1
- [7] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diff-BIR: Towards blind image restoration with generative diffusion prior. In *ECCV*, 2024. 1, 3, 4, 5
- [8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2
- [9] Alireza Shafaei, James J. Little, and Mark Schmidt. AutoRetouch: Automatic professional face retouching. In *WACV*, 2021. 4
- [10] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 1, 2
- [11] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 4
- [12] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 1, 2
- [13] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 4, 5