NTIRE 2024 Restore Any Image Model (RAIM) in the Wild Challenge: Supplementary Material

Jie Liang Radu Timofte Qiaosi Yi Shuaizheng Liu Lingchen Sun
Rongyuan Wu Xindong Zhang Hui Zeng Lei Zhang Yibin Huang Shuai Liu
Yongqiang Li Chaoyu Feng Xiaotao Wang Lei Lei Yuxiang Chen
Xiangyu Chen Qiubo Chen Fengyu Sun Mengying Cui Jiaxu Chen Zhenyu Hu
Jingyun Liu Wenzhuo Ma Ce Wang Hanyou Zheng Wanjie Sun
Zhenzhong Chen Ziwei Luo Fredrik K. Gustafsson Zheng Zhao Jens Sjölund
Thomas B. Schön Xiong Dun Pengzhou Ji Yujie Xing Xuquan Wang
Zhanshan Wang Xinbin Cheng Jun Xiao Chenhang He Xiuyuan Wang
Zhi-Song Liu Zimeng Miao Zhicun Yin Ming Liu Wangmeng Zuo Shuai Li

1. Teams and Methods

In this supplementary material, we briefly describe the participating teams and their proposed methods in the NTIRE 2024 Restore Any Image Model (RAIM) in the Wild Challenge.

1.1. Team MiAlgo

Team MiAlgo proposed a Wavelet UNet with a Hybrid Transformer and CNN model optimized by adversarial training to tackle the real-world image restoration task.

1.1.1 Generator model

As shown in Fig. 1, the model is based on the MWRCAN [8]. The model uses a UNet architecture that employs Harr wavelet transforms and inverse transforms for 2× down-sampling and upsampling. The major convolution modules consist of N Resblocks, where N is 8 in this case. The channels of the Resblocks are marked in the diagram, and there is also a residual connection in each downsample or upsample block, they omitted these connections for the sake of diagrammatical clarity.

Self-attention in transformers enables the network to identify self-similar features throughout the entire image, thereby enhancing its semantic recognition capabilities. However, the attention structure becomes increasingly time-consuming as the feature size grows, rendering it impractical for high-resolution image restoration tasks. To strike a balance between performance and efficiency, the team integrated RESATT structures into the middle block of UNet. RESATT comprises N basic blocks, each consisting of a res-block followed by a single-head self-attention block.

The UNET produces a 3-channel image called out1. To enhance the quality of the restored image, they incorporate a refinement module based on the EMVD[16] approach. This module helps to recover important details that may have been lost during the restoration process. The refinement module takes in the LR image and out1 as inputs and produces a single-channel fusion weight, denoted by α. The final output image is obtained by blending the LR image and out1 using α, i.e., $HR = (1 - \alpha)LR + \alpha out1$. The refinement module is lightweight, comprising only three convolutional layers with a maximum of 16 channels. Despite its simplicity, it is capable of capturing details that are crucial for the final output image.

The team still insists on using GAN models for general restoration because they have found that diffusion models can lead to unacceptable distortions in text and regular textures. The model has approximately 341MB parameters and takes up 7GB of GPU memory and 180ms to infer a 512×512×3 image on a computer with a 4090GPU.

1.1.2 Image degradation

The official competition only provided 100 pairs of training data, as well as 200 images without ground truth in the validation/test phase. They found that the degradation level of the provided 100 pairs of training data is only consistent with 100 images in phase 2, which is relatively mild. The other images in Phase 2 and Phase 3 have a heavier blurring.

Based on the analysis presented, the team developed two GAN degradation models that introduce varying levels of blurring. They enlarge the generator in [19] by doubling the channels, as the degradation model. The first model was trained with the ESRGAN [28] training method and con-

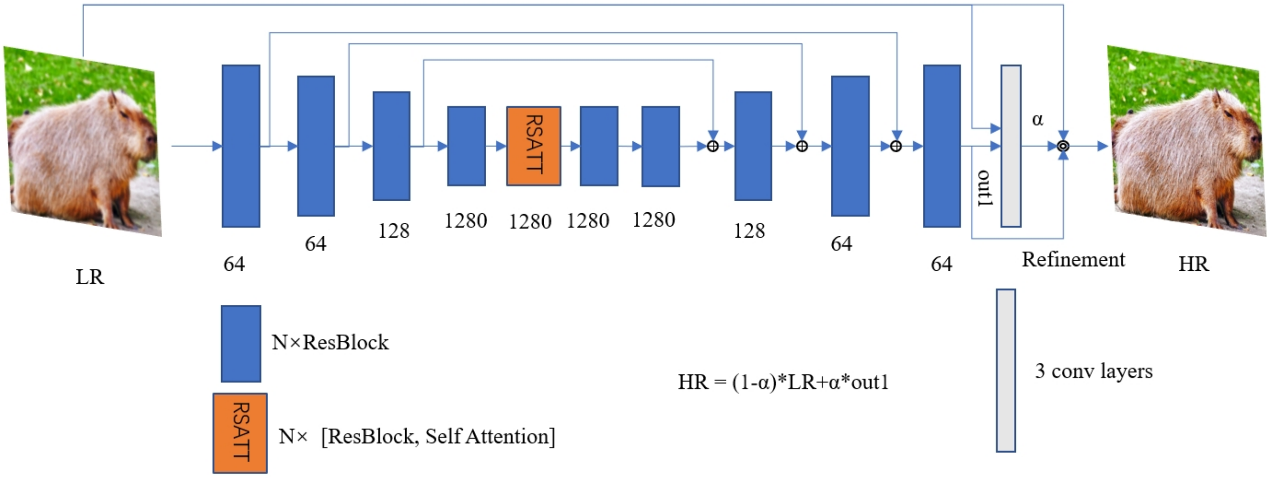


Figure 1. The overall pipeline of the solution proposed by team MiAlgo.

sisted of 100 pairs of training sets, with high-resolution images serving as input to the degradation GAN model and low-resolution images as ground truth. This model introduced a weak level of blurring.

For the second model, they fine-tuned the weak degradation model using the approach outlined in Ref [19]. They trained this model in an unpaired manner, using 50 high-blurring images from phase 2 as unpaired GT and 1000 high-resolution input images from similar scenes as unpaired input. This model introduced a higher level of blurring compared to the first model. When using the second degradation model, they utilize a human segmentation model and a text segmentation model to segment out the human images with heights ¡300 pixels and the text with heights ¡50 pixels. These segments are then replaced with the degradation results from the first degradation model. This strategy helps to reduce the gap between the input and ground truth for small human images and text, and the team has found that this trick improves the fidelity of the results in these regions.

1.1.3 GAN training

The team has an internal ultra-high-definition dataset consisting of approximately 10,000 images. The main scenes include common animals and plants, Chinese and English text, as well as some common urban and rural scenes, which can cover the typical shooting scenarios of mobile phones. They used the two aforementioned degraded GAN models to degrade these images, resulting in a dataset of 20,000 training pairs.

To develop a high-quality image restoration model for phase 2 quantitative measures, they utilized a GAN model trained on 10,000 degraded training pairs from the initial degradation model. The Generator's learning rate was set to 1e-5, with a batch size of 24 and a patch size of 512. The team began training with only L2 loss for $\sim$10,000 iterations, then fixed the loss to include $L2 + 1 * PerceptualLoss + 0.1 * GANLoss$ for an additional 140,000 iterations. They then fine-tuned the model for $\sim$20,000 iterations with $L2+0.1*PerceptualLoss+0.01*GANLoss + 4 * LPIPS$ and a lower learning rate of 1e-6 on the official training set (100 pairs) to achieve a slightly higher quantitative score. The discriminator setting is the same as RealESRGAN [27].

For phase 3, the team continued fine-tuning the model for approximately 100,000 iterations using $loss = L2 + 0.1 * PerceptualLoss + 0.01 * GANLoss + 4 * LPIPS$, with a learning rate of $1e - 5$. They used a mixed dataset with 80% strong degradation and 20% weak degradation by adjusting the training file list ratio. Finally, they crop each training image into 512×512 patches and select the top 10 patches with the higher NIQE score for each image. They continued fine-tuning the model on this subset with a learning rate of $1e - 6$ for $\sim$50,000 iterations. Higher NIQE patches generally have richer textures and they found that fine-tuning the model on this subset resulted in better image details.

1.2. Team Xhs-IAG

Team Xhs-IAG proposed method by combining SUPIR and DeSRRA, which achieves good generative performance and simultaneously acceptable stability on fidelity.

1.2.1 Detailed Method Description for Phase2

```
        window_size = 32,
        embed_dim=180,
        depths=(6, 6, 6, 6, 6, 6),
        num_heads=(6, 6, 6, 6, 6, 6),
        mlp_ratio=4.,
```

The dataset they used is LSDIR[11]. During training, they construct pairs with a resolution of 128x128. The degradation hyperparameters are the same as those for real-esrgan. They trained 92k iterations with batch size=12 (3 for one GPU, total 4 GPUs) in stage-1 and Adam's learning rate is 1e-4.

In the second stage of training, the team added adversarial loss and perceptual loss, and instead of using lsdir, they **only** used 100 paired images provided by the official competition. The results show that the degradation distribution of the official evaluation data is close to that of the 100 images. The specific loss function coefficients are shown below. They trained for a total of 140k iterations in the second stage, with a batch size of 12. The learning rate for Adam is 5e-5.

```
    discriminator=dict(
        type='UNetDiscriminatorWithSpectralNorm',
        in_channels=3,
        mid_channels=64,
        skip_connection=True),
    pixel_loss=dict(type='L1Loss', loss_weight=1.0, reduction='mean'),
    perceptual_loss=dict(
        type='PerceptualLoss',
        layer_weights={
            '2': 0.1,
            '7': 0.1,
            '16': 1.0,
            '25': 1.0,
            '34': 1.0,
        },
        vgg_type='vgg19',
        perceptual_weight=1.0,
        style_weight=0,
        norm_img=False),
    gan_loss=dict(
        type='GANLoss',
        gan_type='vanilla',
        loss_weight=5e-2,
        real_label_val=1.0,
        fake_label_val=0),
```

There is nothing special about the test. For an image, just input it directly into the trained model.

1.2.2 Overall Approach

In recent years, the diffusion method has achieved remarkable results in the field of image generation, and many methods have recently explored its application in the field of image restoration. Due to the unavailability of data for phase 3 of this competition, the distribution of degradation may differ from phase 2. To increase the generalization ability of our solution, the team use SUPIR[35] as our baseline model.

SUPIR is trained on 20 million images and has good modeling of the distribution of natural images. It supports multiple parameters such as positive prompt, negative prompt, and Classifier free guidance scale to adjust the enhanced results. Due to the short competition time and the lack of open-source training code for SUPIR, they did not perform any training fine-tuning on SUPIR, but based on its RGB results. To obtain preliminary RGB results, most official default configurations have not been changed. Only the parameters listed in the table 1 are different from the default parameters.

Although the results generated by diffusion can be natural in most scenes, fidelity issues may arise in some small texture scenes, such as text, patterns, and architectural lines. Especially in the field of photography, this distortion may be unacceptable to professionals, and even worse than not being processed. To alleviate this issue, as shown in figure 2, they will perform another fusion process based on the SUPIR results to obtain the final result. The input of the fusion module includes the SUPIR result, the original image, and a 0/1 mask. To obtain this 0/1 mask, they used the DeSRA[33] method. For the sake of fidelity, the fusion module will perform a lighter enhancement on the area with a value of 1 in the mask (e.g., using GAN-based methods), while the area with a value of 0 will be kept as unchanged as possible(i.e., using SUPIR's result). They introduce our fusion module and DeSRA method in sections 1.2.3 and 1.2.6 in detail, respectively.

Table 1. Modified config parameters for SUPIR inference

config	**default**	**ours**
positive prompt	Cinematic, High Contrast, highly detailed, taken using a Canon EOS R camera, hyperdetailed photo-realistic maximum detail, 32k, Color Grading, ultra HD, Extreme meticulous detailing, skin pore detailing, hyper sharpness, perfect without deformations.	Cinematic, High Contrast, highly detailed, taken using a Canon EOS R camera, hyperdetailed photo-realistic maximum detail, 32k, Color Grading, ultra HD, extremely meticulous detailing, skin pore detailing, hyper sharpness, perfect without deformations, **window glass is very clean**
edm_steps	50	100
sdxl_ckpt	sd_xl_base_1.0_0.9vae	Juggernaut-XL_v9_RunDiffusionPhoto_v2
s_cfg	4.0	2.0

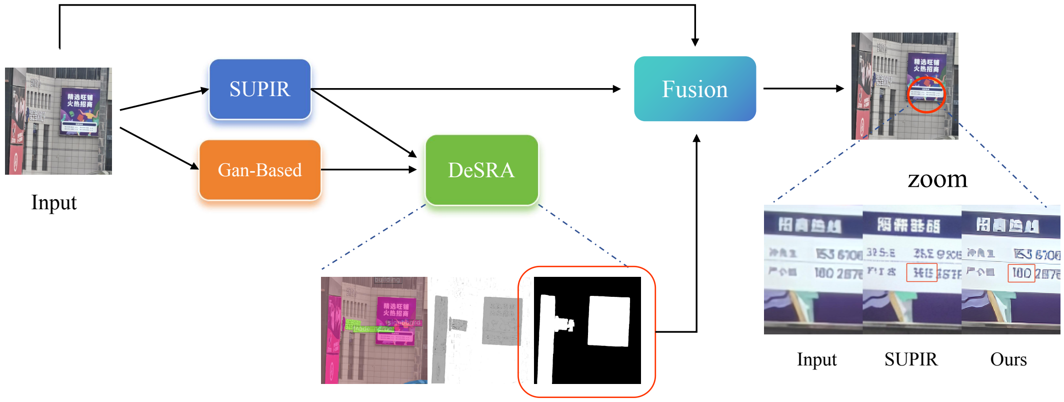


Figure 2. Overall Pipeline of the solution of team Xhs-IAG.

1.2.3 Fusion Network

1.2.4 Architecture

To ensure the authenticity of the results from diffusion-based models, their fusion module performs fine-tuning based on a binary mask. Specifically, the model takes in three components **during inference**: the output from SUPIR, the original image, and a binary mask. Areas, where the mask is zero, indicate that the results from SUPIR are already optimal and do not necessitate any modifications, so they will keep this area. Conversely, regions where the mask is one suggest that the results require re-generation to maintain fidelity. They will replace this area with the corresponding LR part to input the model.

In light of the above, the fusion module operates akin to an image inpainting task[38, 39], with the key difference that the masked areas are not entirely devoid of information; instead, they contain low-quality images that are awaiting enhancement. In the training process, the team continue to follow the Real-ESRGAN strategy to generate paired (LR, GT) on the LSDIR dataset. As illustrated in the figure3, their model backbone continues to employ SRFormer[43] (consistent with Phase 2), with the only change being the inputs. At this point, the input will encompass the LR, mask, as well as the GT and LR combinations derived from the mask. In the inference process, the GT depicted in Figure3 should be substituted with the outcomes yielded by SUPIR.

For the mask used during training, they generate it randomly following the method outlined in STTN[38], while during testing, they utilize the DeSRA[33] approach to obtain the mask. Regarding the DeSRA method, it will be introduced later.

Given that the inputs already contain regions of high quality, the loss function must be correspondingly modified to account for this. Similar to image inpainting tasks[39], the loss function encompasses hole loss, valid loss, perceptual loss, and adversarial loss. Notably, for the generated fake image, the discriminator employs the technique of using soft labels when calculating the least square loss[39], rather than the hard labels of 0 and 1. This design allows the discriminator to better discern potential mask areas.

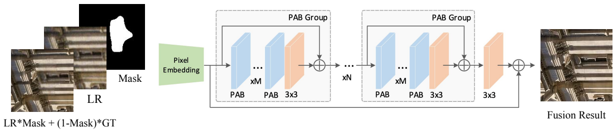


Figure 3. Fusion module of the solution of team Xhs-IAG.

1.2.5 Training Details

The team used SrFormer[43] as the backbone, the specific parameters are shown in the following code.

```
    window_size = 24,
    embed_dim = 360,
    depths=(6, 6, 6, 6, 6),
    num_heads=(6, 6, 6, 6, 6),
    mlp_ratio=3
```

The dataset they used is LSDIR[11]. During training, the team constructed pairs with a resolution of 144x144. The degradation hyperparameters are the same as those for real-esrgan. They trained 172k iterations with batch size=8 (2 for one GPU, total 4 GPUs) and Adam's learning rate is 1e-4. The specific loss function coefficients are shown below.

```
    valid_loss = dict(type='Valid_loss',
        loss_weight=0.3),
    hole_loss = dict(type='Hole_loss',
        loss_weight=0.01),
    perceptual_loss=dict(
        type='PerceptualLoss',
        vgg_type='vgg19',
        layer_weights={
            '1': 1.,
            '6': 1.,
            '11': 1.,
            '20': 1.,
            '29': 1.,
        },
        layer_weights_style={
            '8': 1.,
            '17': 1.,
            '26': 1.,
            '31': 1.,
        },
        perceptual_weight=0.2,
        style_weight=150,
        norm_img=False,
    ),
    gan_loss=dict(
        type='GANLoss',
        gan_type='lsgan',
        loss_weight=0.02,
        real_label_val=1.0,
        fake_label_val=0)
```

The generation of random masks during training can be referenced at the specified line in the following GitHub repository: STTN GitHub Repository.

1.2.6 DeSRA Method

With the fusion model in place, it is necessary to ascertain the masks used during testing, specifically identifying the regions where the diffusion results are distorted. A straightforward method involves manual annotation of masks, but this approach is not only unfair in the context of competition but also labor-intensive.

The team employ the methodology from DeSRA[33] for identifying GAN artifacts, utilizing a combination of structural similarity metrics and semantic segmentation outcomes to generate masks. To be precise, they ascertain the mask by contrasting the outputs from the GAN model with those from the diffusion model. The GAN model utilized in this process is the one that has been adequately trained during Phase 2. This choice is motivated by the fact that, despite the GAN model's potential shortcomings in visual quality, it excels in preserving fidelity in intricate details such as text and textures. By adjusting the parameters, the team strives to align the distribution of the masks with human visual perception. It is important to note that no special parameters are used for any individual image; the same set of parameters is applied consistently across all 50 images.

To enhance the accuracy of the segmentation, they have utilized the Mask2Former model[6] for this task. Compared to the SegFormer model[32] used in the original DeSRA, Mask2Former represents a more advanced approach. Within the provided code, they have included scripts for mask generation, which encompass all the parameters used, including the weights for semantic categories, contrast_threshold, area_threshold, and so on.

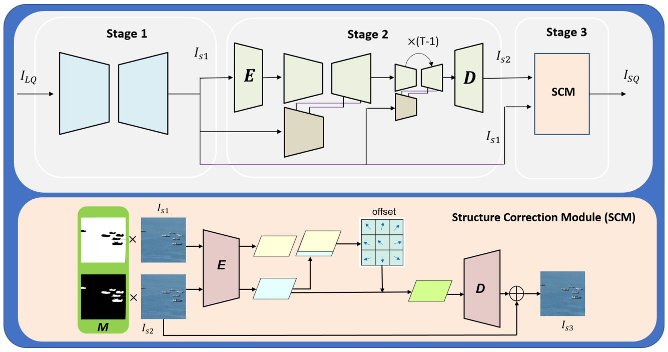


Figure 4. The three-stage pipeline of CGSD proposed by team So Elegant.

1.3. Team So Elegant

The team proposed a Consistency-guided Stable Diffusion method for Image Restoration.

As shown in Figure 4, the proposed Consistency Guided Stable Diffusion (CGSD) model has three primary stages. Stage 1 is based on a CNN-based restoration model DiffIR [31] to remove diversified degradations. DiffIR uses the powerful mapping capability of the diffusion model to estimate a compact IR prior representation (IPR) to guide image restoration, thereby improving the recovery efficiency and stability of the diffusion model in image restoration. To bridge the domain gap, the degradation of the given data is used to customize the degradation distribution for training [42], which improves the performance of the target test images while maintaining generalization performance. Additionally, BSRGAN [40] is used to simulate image degradation to generate pairs of data for training. And, virtual focus blur is added to BSRGAN to better suit the target test images. For stage 2, Stable Diffusion (SD) [21] is leveraged to refine the texture and details. To improve the fidelity of SD model restoration, a Consistency-Guided Sampling Module (CGS) is proposed to limit the generation. Specifically, the CGS module takes the recovered image of stage 1 as the consistent guidance in each decoding step and aligns the recovery results of each step with it:

$$x_{t-1} \leftarrow x_{t-1} + \sigma_t (x_{s1} - x_{t-1}) \tag{1}$$

where x_{t-1} and x_{s1} corresponds to the noise-free predicted output at step $t-1$ and the recovered I_{s1} latent. σ_t represents the weight of the guidance. The image structure is determined in the early diffusion step, and the later stage mainly generates high-frequency details. The final stage 3 is proposed to address the contexture distortion caused by the diffusion model. The contextual information from I_{s1} guides the refined image I_{s2}. Similar to [10], deformable convolution[7] is employed to warp the details in I_{s2} to match the fidelity of I_{s1}. A problematic mask [33] M

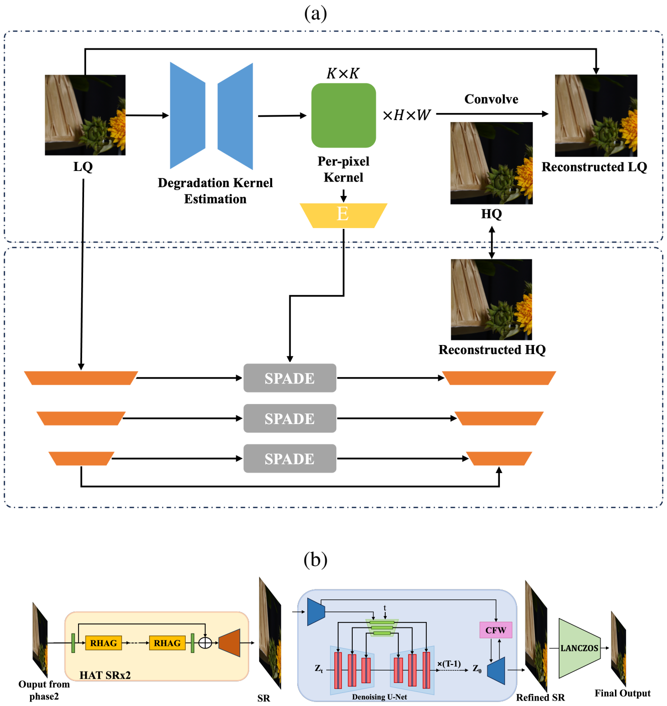


Figure 5. A visual representation of the solution proposed by IIP_IR.

located by a relative local variance distance from I_{s1} and I_{s2} and semantic-aware thresholds are used as the additional condition. The method is implemented in Pytorch and trained using 8x Nvidia V100 GPU for training. For stage 1, the team first uses the original configuration from DiffIR for training and then adjusts the learning rate to 5e-5, batch size to 2, and trains 10K iterations at a resolution of 512x512. For stage 2, they train the SD model using the AdamW [13] optimizer with a learning rate of 1e-4 and a batch size of 64 for 50K steps. For stage 3, they use a batch size of 2 and a patch size of 1024x1024 for training. Adam is used as the optimizer with a learning rate of 1e-4. And they train the model for 20K iterations.

1.4. Team IIP_IR

The team IIP_IR has introduced an integrated framework called Degradation-Aware Image Restoration(DAIR) based on the FFTformer architecture introduced in [9] for phase 2. DAIR comprises three main components: Degradation Kernel Estimation (DKE), Degradation Representation Injection (DRI), and FFTfromer. The team's innovative approach, as illustrated in Figure 5a, has the potential to enhance existing models and improve overall performance.

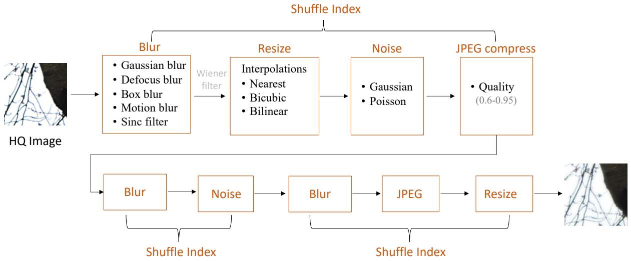


Figure 6. Overview of the synthetic low-quality image generation proposed by team DACLIP-IR.

To enable the model to process the degradation of the images, they utilize a method to learn per-pixel degradation convolution kernels similar to blur kernels, which can reconstruct LQ when convolved with HQ images. Unlike the blur kernel, DKE does not constrain the reconstructed kernel to have positive weights that sum to one, thus learning richer degenerate representation.

To maximize the retention of degraded information for image restoration models, the kernels estimated by DKE will be embedded into the Spatially Adaptive representation and injected into U-Net architecture, which is processed through a SPADE module [20]. The processing of the SPADE module does not change the network structure, thus DKE and DRI can be applied directly to any Unet-based image restoration model.

In the training process, the team uses the method mentioned in [4] to generate paired data for pre-training the model, improves its generalization ability and adaptability, and finally fine-tunes the model using 100 pairs. While L1 loss normally trained networks which usually produce smooth/blurry results, they apply perception loss and GAN loss constraints to reconstructed LQ and HQ for both the pre-training phase and fine-tuning phase to increase the realism of the image.

Figure 5b illustrates the pipeline of phase 3. The team utilizes the model to refine the details of the pre-processed images from phase 2. The images first undergo x2 upscaling using HAT[5] to enrich the textures. The initial upscaling phase effectively mitigates distortions of small-scale details such as texts during the texture generation process leveraging pre-trained diffusion priors. They employ StableSR[25] with SD-Turbo, to further refine the upscaled images, producing realistic textures in regions with severe degradations. The refined images were then downscaled with LANCZOS interpolation to obtain the final output.

1.5. Team DACLIP-IR

Team DACLIP-IR proposed a photo-realistic image restoration method with enriched vision-language features.

The model is built upon the IR-SDE [15] and DACLIP-UIR [14]. Since no training datasets are provided in this challenge, the team chooses to generate LQ images using a similar pipeline as in Real-ESRGAN [29] but with an index-shuffling strategy, as shown in Figure 6. Based on the synthetic dataset, they retrain DA-CLIP to enhance LQ features by minimizing an ℓ_1 distance between LQ embeddings and HQ embeddings. Then they incorporate the enhanced LQ embeddings into IR-SDE with cross-attention to restore HQ images, similar to DA-CLIP [14]. In addition, they propose a posterior sampling approach for IR-SDE that improves both fidelity and perceptual performance. To further improve the generalization ability, they first train the model on the LSDIR dataset [11] and then finetune it on a mixed dataset with both synthetic and real-world images for phase two and phase three. Note that they use the same model for phase two and phase three, but take the original reverse-time SDE for phase three for better visual performance (small noise makes the photo look more realistic).

Specific training details for phase two: The team adds the paired validation dataset in phase one to further finetune the model, which improves a lot across all metrics.

Specific training details for phase three: They use the same model trained from phase two for phase three. To make the image look non-smooth and oil-painted, they use the original reverse-time SDE during inference.

1.6. Team TongJi-IPOE

Team TongJi-IPOE proposed a DRBFormer-StableSR fusion Network for restoring any image model in the Wild.

Method. The overall architecture is shown in Figure 7. The proposed network consists of two parts: DRBFormer image restoration network and StableSR [26] image SR network. DRBFormer uses Restormer Blocks as the backbone. Inspired by [36], a multi-scale dynamic residual module DRB is designed in the decoding network to better to better handle the varying blur [23]. Considering that Diffusion priors can improve the performance of restored images, the network adopts the fusion method of Eq. (2) for image restoration. Due to the randomness of the diffusion model, the generated image may deviate from the real situation, so the adjustable coefficient t was set to 0.9 in this competition.

$$\hat{I} = t * DRBFormer(I_{blur}) + (1 - t) * StableSR(I_{blur}) \tag{2}$$

where $\hat{I}$ is the result of restoration, $t \in [0, 1]$ is adjustable coefficient and I_{blur} is blurryimage.

Training strategy. In total, four datasets are used including DPDD[2], SIDD[1], GoPro[17] and NH-HAZE[3]. To train the models with images, the training dataset is augmented with random clipping. The details of the training steps are as follows:

1. Pretraining on combined datasets. Ground truth patches of size 128×128 are randomly cropped from Ground truth images, and the mini-batch size is set to 8. The model is trained by minimizing weighted L1 loss and perceptual loss function with Adam optimizer. The initial learning rate is set to $3\times10^{-}4$ and the total number of iterations is 392k.

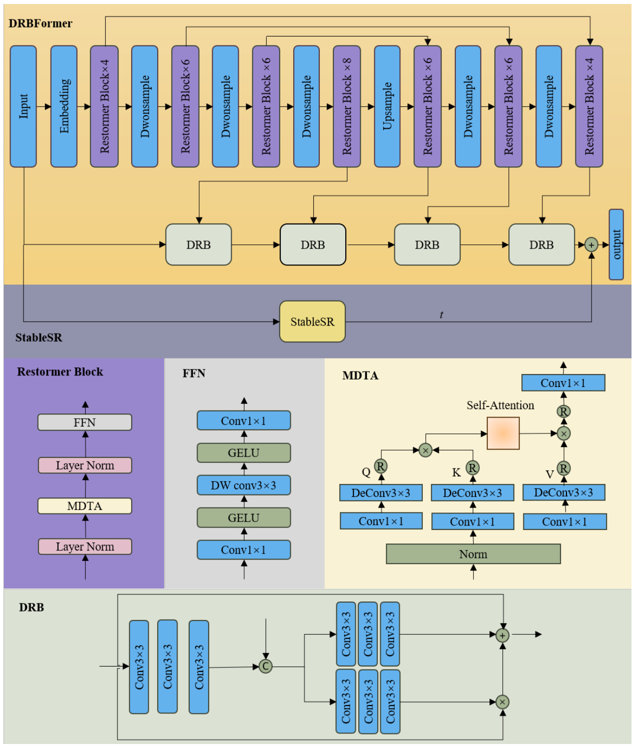


Figure 7. The overall architecture of the proposed method from team TongJi-IPOE.

2. Finetuning on combined datasets. For the model to adapt to higher resolution image processing, crop the image to 160×160,192×192,256×256,320×320,384×384, and set the mini-batch size to [5,3,2,1,1]. The model is trained by minimizing weighted L1 loss and perceptual loss function with Adam optimizer. The initial learning rate is set to $3\times10^{-}4$ and adjusted by cosine annealing. The total number of iterations is 208k.

1.7. Team ImagePhoneix

Team ImagePhoneix adopted DiffIR [31] as the baseline network, as shown in Figure 1. They froze the “stage 2” of the DiffIR and fine-tune its “stage 1” network on the provided LR-HR image pairs.

Implementation details.

With provided image pairs, they first cropped them into sub-images of the size 400×400 for accelerating I/O speed, resulting in a total number of 2500 sub-images. To fine-tune the pre-trained model, all the sub-images are cropped into image patches with the size 256×256. They randomly flipped and rotated the input images for data augmentation.

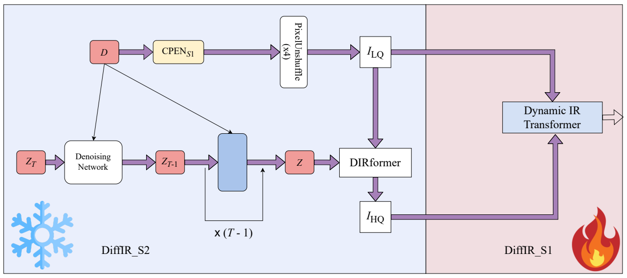


Figure 8. The technical pipeline adopted by the team Image-Phoneix.

Adam algorithm is adopted with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to update the model parameters. They set the initial learning rate and the total number of iterations to 1×10^{-4} and 1×10^5, respectively. However, the encoder is updated with a different strategy, which updates the model parameters of the encoder in 2.5×10^4 iterations and sets the initial learning rate to 2×10^{-4}. The learning rate of the encoder is decay with a factor of 0.1 in the 1.5×10^4-th iteration. Different from the encoder, the learning rate of the image generator is decay with a factor of 0.5 at the 8.0×10^4-th iteration.

In Phase II, the evaluation metric is a linear combination of the reconstruction and perceptual measurements. To handle this issue, The team adopted a hybrid loss function to fine-tune the model, which involves L_1 loss, perceptual loss based on VGG features L_{vgg}, adversarial loss L_{GAN}, and Kullback–Leibler divergence L_{KL}. The total loss is defined as $\mathcal{L} = \lambda_1 L_1 + \lambda_2 L_{\text{vgg}} + \lambda_3 L_{\text{GAN}} + \lambda_4 L_{\text{KL}}$, where L_1 loss measure the reconstruction error of the generated images, L_{vgg} aims to improve the perceptual quality of images, L_{GAN} and L_{KL} measure the distribution distance between the generated images and the ground-truth images in the spatial and latent spaces, respectively. $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are hyper-parameters to balance the distortion and perceptual quality of images and set to 1.0 in this Phase.

1.7.1 Phase III: Evaluation on Subjective Measurements

In Phase III, the team aims to improve the perceptual quality of generated images. Instead of using perceptual loss based on the VGG features, they adopt the robust distribution loss [18] which minimizes the distribution distance between the generated images and the ground-truth images based on Fast Fourier transform (FFT). Given the generated image x and the ground-truth image y, the robust distribution loss L_{freq} is defined as follows:

$$L_{\text{freq}}(x, y) = L_{\text{WD}}(\mathcal{A}_x, \mathcal{A}_y) + \lambda_{\text{phase}} L_{\text{WD}}(\mathcal{P}_x, \mathcal{P}_y), \quad (3)$$

where $\mathcal{A}_x = |\mathcal{F}(x)|$ and $\mathcal{A}_y = |\mathcal{F}(y)|$ denote the frequency spectrum of the images x and y via FFT $\mathcal{F}$, respectively. $\mathcal{P}_x$ and $\mathcal{P}_y$ represent the phase of $\mathcal{F}(x)$ and $\mathcal{F}(y)$, respectively. L_{WD} is the Wasserstein distance, and λ_{phase} is the hyper-parameter that is set to 0.1 in the fine-tuning procedure.

1.8. Team HIT-IIL

The team HIT-IIL used the degradation process of Real-ESRGAN [29] and replaced the backbone with Restormer [37]. For phase 2, they only trained a Real-ESRGAN x1plus model with an additional lpips loss. For phase 3, they used the backbone of Restormer to train a new x1model and averaged the results with weights 0.8 and 0.2, respectively.

They use DF2K (DIV2K and Flickr2K) datasets to train the model. For pre-processing, they use a multi-scale strategy, i.e., they downsample HR images to obtain several Ground-Truth images with different scales. They then crop DF2K images into sub-images for faster IO and processing.

1.9. Team MARSHAL

1.9.1 Methods details

The team observed that the input images and evaluation criteria of the two phases are different. The input images in phase 2 have higher quality. The evaluation criteria for this phase are based on reference evaluation indicators. The input quality of phase 3 is relatively low, with a more serious blur. This phase uses the method of manual scoring to select images with better visual effects as the winners. Taking into account the existing solutions, the team decided to adopt a gan-based approach in phase 2 to obtain higher objective indicator scores. In phase 3, a diffusion-based approach is adopted to make the results more visually appealing.

1.9.2 Phase 2

The organizers provided 100 pairs of training images whose input quality and imaging style are similar to the test set of the first stage. Therefore, they chose DiffIR [31] for this stage. It only uses the diffusion process to model the condition branch, and the main network is trained using the GAN loss, so it rarely destroys local details (such as text, small faces), and can obtain a higher objective evaluation index. They directly use the pre-trained model of DiffIR and fine-tune it with the paired dataset provided by the organizer, so that the team can quickly obtain a good result. The whole finetuning process continues 6.6 k iterations with a batch size of 48. In addition, in the process of preparing the dataset, they adopted a multi-scale downsampling strategy, hoping that the model could gain knowledge of different scales. The downsampling scales are set to 0.75, 0.5, and 0.33, respectively.

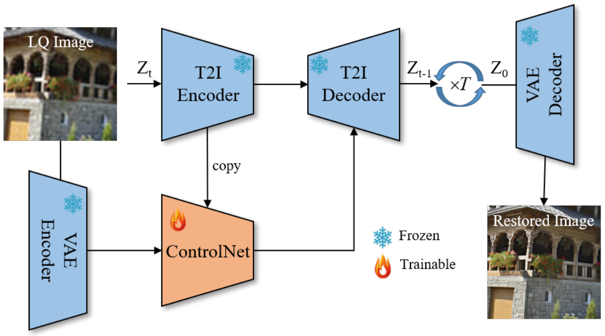
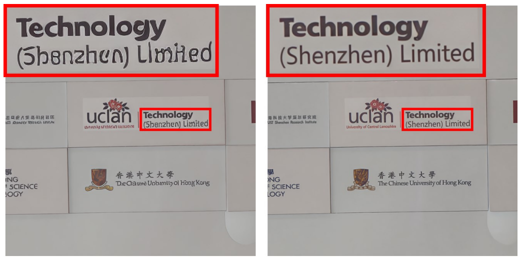


Figure 9. The pipeline of the solution proposed by team MARSHAL.



w/o resize

w resize

Figure 10. Comparison of resize strategies on small text scenarios in the solution proposed by team MARSHAL.

1.9.3 Phase 3

In phase 3, the test set provided by the organizer has a significant domain gap from the test set of phase 2, and the degradation is more severe. The team thinks they cannot directly use the model that performs well in phase 2 to obtain good visual results in the second stage, so they switched to using the methods [24, 25, 30, 34] of the pre-training diffusion model. As shown in Fig. 9, the team chooses the popular ControlNet [41] as the solution. Following [12], they use pretrained VAE encoder as the image encoder. In terms of training data, they choose LSDIR [11], which contains tens of thousands of texture-rich images. As for data degradation, to match the more severe degradation of the test set, they choose realesrgan's [29] degradation pipeline to synthesize paired data. They train the model with a batch size of 32 for 100k iterations. In the inference stage, the team resizes the input to 2048 before feeding it into the model, which aims to preserve small structures like texts as shown in Fig. 10. The team also adopts the LRE strategy proposed in [30] to improve fidelity. The pre-trained diffusion model in this solution is SD2-base [22].

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018.

[2] Abdullah Abuolaim and M. S. Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, 2020.

[3] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1798–1805, 2020.

[4] Guillermo Carbajal, Patricia Vitoria, Mauricio Delbracio, Pablo Musé, and José Lezama. Non-uniform blur kernel estimation via adaptive basis decomposition. *arXiv preprint arXiv:2102.01026*, 2021.

[5] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023.

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[8] Andrey Ignatov, Radu Timofte, Zhilu Zhang, Ming Liu, Haolin Wang, Wangmeng Zuo, Jiawei Zhang, Ruimao Zhang, Zhanglin Peng, Sijie Ren, et al. Aim 2020 challenge on learned image signal processing pipeline. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 152–170. Springer, 2020.

[9] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023.

[10] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European conference on computer vision*, pages 399–415. Springer, 2020.

[11] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023.

[12] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[13] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[14] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 2023.

[15] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023.

[16] Matteo Maggioni, Yibin Huang, Cheng Li, Shuai Xiao, Zhongqian Fu, and Fenglong Song. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3466–3475, 2021.

[17] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 257–265, 2016.

[18] Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, and Lin Ma. Misalignment-robust frequency distribution loss for image transformation. *arXiv preprint arXiv:2402.18192*, 2024.

[19] Qian Ning, Jingzhu Tang, Fangfang Wu, Weisheng Dong, Xin Li, and Guangming Shi. Learning degradation uncertainty for unsupervised real-world image super-resolution. In *IJCAI*, pages 1261–1267, 2022.

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[23] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16283–16292, 2022.

[24] Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023.

[25] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023.

[26] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *ArXiv*, abs/2305.07015, 2023.

[27] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*.

[28] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[29] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

[30] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. *arXiv preprint arXiv:2311.16518*, 2023.

[31] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023.

[32] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021.

[33] Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: detect and delete the artifacts of gan-based real-world super-resolution models. *arXiv preprint arXiv:2307.02457*, 2023.

[34] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023.

[35] 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. *arXiv preprint arXiv:2401.13627*, 2024.

[36] Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2021.

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022.

[38] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting.

In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 528–543. Springer, 2020.

[39] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[40] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021.

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[42] Ruofan Zhang, Jinjin Gu, Haoyu Chen, Chao Dong, Yulun Zhang, and Wenming Yang. Crafting training degradation distribution for the accuracy-generalization trade-off in real-world super-resolution. In *International Conference on Machine Learning*, pages 41078–41091. PMLR, 2023.

[43] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12780–12791, 2023.