2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-0249-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPRW59228.2023.00169

Refusion: Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion Models

Ziwei Luo Fredrik K. Gustafsson Zheng Zhao Jens Sjölund Thomas B. Schön
Uppsala University, Sweden
{ziwei.luo,fredrik.gustafsson,zheng.zhao,jens.sjolund,thomas.schon}@it.uu.se

https://github.com/Algolzw/image-restoration-sde

(1) Shadow Removal *(2) Stereo Super-Resolution* *(3) Bokeh Effect Transform* *(4) HR Dehazing*

Figure 1. Visual results of applying our Refusion to various synthetic and real-world restoration tasks, including (1) Shadow Removal, (2) Stereo Super-Resolution, (3) Bokeh Effect Transform, and (4) HR Non-Homogeneous Dehazing. Note that all tasks require processing of large-size images (2K-6K resolution) with complex degradations, which is far beyond the capabilities of existing diffusion models.

Abstract

This work aims to improve the applicability of diffusion models in realistic image restoration. Specifically, we enhance the diffusion model in several aspects such as network architecture, noise level, denoising steps, training image size, and optimizer/scheduler. We show that tuning these hyperparameters allows us to achieve better performance on both distortion and perceptual scores. We also propose a U-Net based latent diffusion model which performs diffusion in a low-resolution latent space while preserving high-resolution information from the original input for the decoding process. Compared to the previous latent-diffusion model which trains a VAE-GAN to compress the image, our proposed U-Net compression strategy is significantly more stable and can recover highly accurate images without relying on adversarial optimization. Importantly, these modifications allow us to apply diffusion models to various image restoration tasks, including real-world shadow removal, HR non-homogeneous dehazing, stereo super-resolution, and bokeh effect transformation. By simply replacing the datasets and slightly changing the noise network, our model, named ***Refusion****, is able to deal with large-size images (e.g.,* $6000 \times 4000 \times 3$ *in HR dehazing) and produces good results on all the above restoration problems. Our Refusion achieves the best perceptual performance in the NTIRE 2023 Image Shadow Removal Challenge and wins* 2^{nd} *place overall.*

1. Introduction

Image restoration is a long-standing problem in computer vision due to its ill-posed nature and extensive demands in industry. Broadly speaking the challenge is to restore the high-quality (HQ) image from the low-quality (LQ) counterpart subject to various degradation factors (e.g., noising, downsampling, and hazing). Over the past decade, methods based on deep learning have achieved impressive performance in image restoration. However, most of these methods are prone to produce over-smooth images due to their pixel-based reconstruction loss functions, i.e.,

L_1/L_2 [20, 50, 51, 58, 76].

Recently, the diffusion model has shown a strong capability in producing high-quality results by sampling images consisting of pure noise and then iteratively denoising them with Langevin dynamics [19, 28, 56] or reverse-time stochastic differential equations (SDEs) [57, 60]. However, many common image restoration tasks (e.g., deraining, dehazing, and deblurring) are still challenging for diffusion models, due to the complex degradations and the large and arbitrary image sizes in real-world datasets. There has been interesting developments when it comes to the use of pre-trained diffusion models. Two drawbacks with the existing approaches are that; 1) they rely on carefully curating the datasets (e.g., ImageNet [18] and FFHQ [31]), 2) they require the degradation parameters to be known. These drawbacks limit their applicability when it comes to real-world tasks [4, 12, 14, 17, 32, 33].

In order to handle intricate real-world distortions, recent developments [47, 54, 69] have utilized a combination of a pure noise image and a low-quality image as an intermediary input for the noise network. This approach avoids the need for degradation parameters and enforces the reverse process to convert the noise into the desired high-quality image. However, these approaches are somewhat heuristic and are difficult to apply to general tasks. A more general image restoration method is IR-SDE [41], which proposes to recover HQ images based on a mean-reverting SDE, which implicitly models the degradation and is applicable to various tasks by changing the datasets only. A drawback with IR-SDE is that it is computationally demanding at test time since it requires multi-step denoising on the full image to restore the final output. This can be problematic for real-world applications, in particular for high-resolution images.

The purpose of this paper is to improve the diffusion model is a way that enhance its effectiveness in tackling diverse real-world image restoration tasks. The result is **Refusion** (*image **Re**storation with diff**usion** models*). Due to its simplicity and flexibility in accommodating different problems, the IR-SDE serves as the foundation for Refusion. By exploring different noise network architectures, we show that using the nonlinear-activation-free network (NAFNet) [7] can achieve good performance in noise/score prediction while at the same time being more computationally efficient. Moreover, we also illustrate the efficacy of different noise levels, denoising steps, training image sizes, and optimizer/scheduler selections. To further deal with large images, we propose a U-Net based latent diffusion strategy. This allows us to perform image restoration in a compressed and low-resolution latent space, which speeds up both the training and the inference. In the experiments, we demonstrate our improved diffusion model on the tasks of real-world shadow removal, HR non-homogeneous dehazing (with images of size $6000 \times 4000 \times 3$), stereo super-resolution, and the bokeh effect transformation. The experiments show that the proposed Refusion is effective on all the image restoration tasks mentioned above.

Our contributions are summarized as follows:

- Compared to existing diffusion-based approaches, our method can gracefully handle high-resolution images by performing image restoration in the U-Net compressed latent space, while preserving high-resolution information from the original input for the decoding process. Importantly, our U-Net compression strategy offers significantly improved stability compared to existing latent diffusion models and can recover high-accuracy images without requiring adversarial optimization.

- We perform a comprehensive empirical study of several factors that have a major impact on the performance of diffusion models for image restoration.

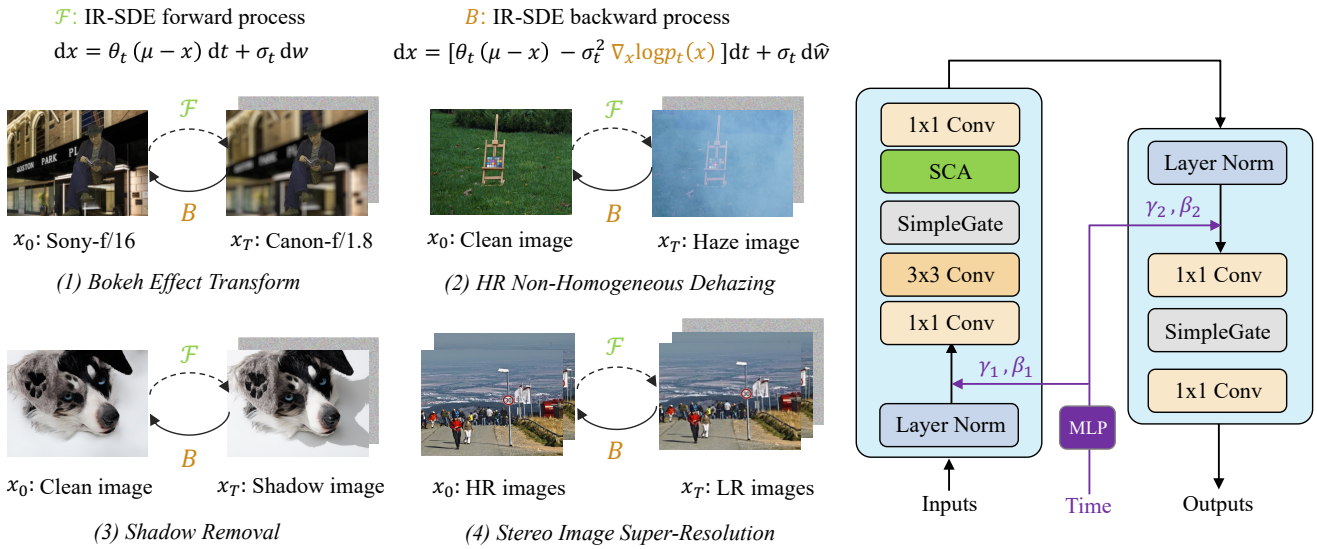
- We propose to change the diffusion base network from U-Net to NAFNet. The latter achieves better image restoration performance across all tasks while requiring fewer model parameters and being computationally more efficient.

- We evaluate our approach on extensive real-world and synthetic datasets, further showing strong versatility to a variety of image restoration problems.

2. Related Work

Image restoration aims to restore a high-quality image from a degraded low-quality version. When it comes to approaches based on deep learning, two early and influential contributions are SRCNN [20] and DnCNN [76]. They made use of convolution neural networks (CNNs) for image super-resolution and denoising, which significantly improved the performance in each application. This development spurred a lot of activity when it comes to making use of CNNs for various image restoration tasks [8, 11, 21, 34, 35, 42, 44, 65, 73–75, 77, 80, 81]. Many of these approaches can be viewed as variations of [20, 76] trained with pixel reconstruction losses such as L_1 and GAN.

Recently, transformer-based architectures [63] have shown impressive performance and hence received a lot of attention when it comes to high-level computer vision tasks [22, 24, 38]. These architectures have also been employed for image restoration [6, 36, 43, 68, 70, 72]. For example, IPT [6] is the first work to propose the use of pre-trained transformers for image processing. Subsequently, SwinIR [36] modifies the Swin Transformer [38] with additional convolution layers and residual connections to achieve state-of-the-art performance on various image



(a) Various image restoration tasks. (b) Modified NAFBlock

Figure 2. (a) Illustrations of various image restoration tasks based on our proposed Refusion method. We use the mean-reverting SDE to recover HQ images from LQ images, adopting the IR-SDE approach [41]. (b) The NAFBlock with an additional time processing branch which is depicted in purple color. Here "SCA" is the simple channel attention, and "SimpleGate" is an element-wise operation that splits feature channels into two parts and then multiplies them as output.

restoration tasks such as image super-resolution and denoising. Restormer [72] and Uformer [68] combine the transformer with U-shape structures to achieve more efficient image restoration. In addition, there are also attempts to make use of the MLP [59] and the nonlinear activation free networks [7] to restore images.

3. Preliminaries: Mean-Reverting SDE

Our method leverages a diffusion model for realistic image restoration. Specifically, we use IR-SDE [41] as the base diffusion framework, which can naturally transform the high-quality image to its degraded low-quality counterpart, irrespective of how complicated the degradation is (even for real-world degradations, see Figure 2). The forward process of the IR-SDE is defined as:

$$dx = \theta_t\,(\mu - x)dt + \sigma_t dw, \tag{1}$$

where θ_t and σ_t are time-dependent positive parameters characterizing the mean-reversion speed and the stochastic volatility, respectively. If we set the SDE coefficients in (1) to satisfy $\sigma_t^2 \,/\, \theta_t = 2\,\lambda^2$ for all times t, the marginal distribution $p_t(x)$ can be computed according to [41]

$$p_t(x) \;= \mathcal{N}\big(x(t) \mid m_t, v_t\big), \tag{2a}$$

$$m_t := \mu + (x(0) - \mu)\,\mathrm{e}^{-\bar{\theta}_t}, \tag{2b}$$

$$v_t := \lambda^2 \left(1 - \mathrm{e}^{-2\,\bar{\theta}_t}\right), \tag{2c}$$

where $\bar{\theta}_t := \int_0^t \theta_z\,\mathrm{d}z$. Note that as t increases, the mean value m_t and the variance v_t converges to μ and λ^2, respectively. Hence, the initial state $x(0)$ is iteratively transformed into μ with additional noise, where the noise level is fixed to λ.

The IR-SDE forward process (1) is a forward-time Itô SDE, which has a reverse-time representation as [57]

$$dx = \big[\theta_t\,(\mu - x) - \sigma_t^2\,\nabla_x \log p_t(x)\big]dt + \sigma_t d\hat{w}. \tag{3}$$

Note that during training we have access to HQ images which means that we can employ (2) to compute the ground truth score function

$$\nabla_x \log p_t(x) = -\frac{x(t) - m_t}{v_t}. \tag{4}$$

The reparameterization trick now allows us to sample $x(t)$ according to $x(t) = m_t(x) + \sqrt{v_t}\,\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, I)$ is a standard Gaussian noise. Then we can rewrite (4) as $\nabla_x \log p_t(x) = -\frac{\epsilon_t}{v_t}$. A CNN network is usually trained to estimate the noise, and at test time we then simulate the backward SDE to transform low-quality images into high-quality versions, similar to other diffusion-based models.

4. Improving the Diffusion Model

We will in Sec. 4.1 introduce the U-Net based latent diffusion model that allows us to perform diffusion in the low-resolution space to significantly improve the sample efficiency. After that, we introduce the nonlinear activation

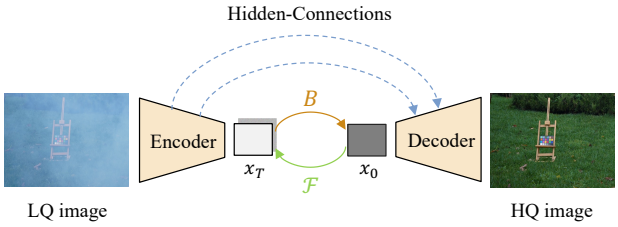


Figure 3. Overview of our U-Net based latent diffusion model. The restoration is performed in a low-resolution latent space.

free blocks (NAFBlocks) [7] to IR-SDE in Sec. 4.2, and outline several training strategies that can improve the restoration performance in Sec. 4.3. As an overview, Figure 2 illustrates the tasks and networks of the proposed Refusion method.

4.1. Latent Diffusion under U-Net

Iteratively running the diffusion model (even with just a few denoising steps) on tasks with high-resolution images is notoriously time-consuming. Especially for HR dehazing, where all images are captured with $6000 \times 4000 \times 3$ pixels, which is far beyond the input size of any existing diffusion model. To handle large input sizes we propose to perform the restoration in a low-resolution latent space, by incorporating a pretrained U-Net network. The overall architecture of the proposed U-Net based latent diffusion model is shown in Figure 3. An encoder compresses the LQ image into a latent representation, which is transformed into an HQ latent representation via the IR-SDE backward process. From this, a decoder then reconstructs an HQ image. An important difference compared to latent-diffusion [52], which uses VAE-GAN as the compressing model, is that the proposed U-Net maintains multi-scale details flowing from the encoder to the decoder through skip-connections. This better captures the input image's information and provides the decoder with additional details to reconstruct more accurate HQ images.

When training the U-Net model, we need to make sure that the compressed latent representation is discriminative and contains the main degradation information. The U-Net decoder must also be able to reconstruct HQ images from transformed LQ latent representations. We therefore adopt a latent-replacing training strategy, as shown in Figure 4. Each LQ image is first encoded and decoded by the U-Net, and a reconstruction L_1 loss is applied. The U-Net is then also trained to reconstruct the corresponding HQ image, by replacing the LQ latent representation with that of the HQ image and running the decoder again. Importantly, our proposed training strategy does not involve any adversarial optimization. The model training is thus more stable than for latent-diffusion [52].

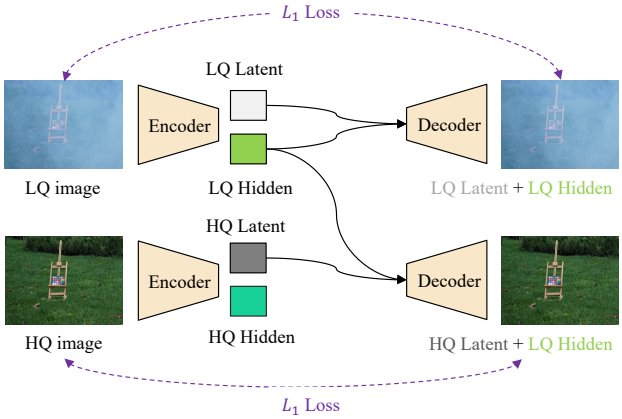


Figure 4. Our proposed latent-replacing pretraining strategy for the U-Net model, utilizing two reconstruction L_1 loss terms.

4.2. Modified NAFBlocks for Noise Prediction

A commonly used architecture for noise/score prediction is the U-Net [53] with residual blocks [25] and attention mechanisms such as channel-attention and self-attention [28, 57]. The recently proposed DiT [49] makes use of a transformer-based structure and simulates diffusion in a lower-resolution latent space, which sets a new state-of-the-art on the class-conditional ImageNet 512×512 generation in terms of FID. But even under the latent-diffusion framework [52], pure transformer architectures still incur a larger computational cost than affordable in traditional image restoration applications.

To address the aforementioned model efficiency problem, we explore a new architecture for noise prediction. Specifically, our noise network is based on slightly modified nonlinear activation free blocks (NAFBlocks) [7]. Nonlinear activation free means that we replace all nonlinear activation functions with the "SimpleGate", an element-wise operation that splits feature channels into two parts and then multiplies them together to produce the output. As illustrated in Figure 2(b), we add an additional multilayer perceptron to process the time embedding to channel-wise scale and shift parameters γ and β, for both the attention layer and feed-forward layer. To adapt to different tasks, we also slightly modify the network with task-specific architectures, such as the lens information in *Bokeh Effect Transform* and dual inputs in *Stereo Image Super-Resolution*. The learning curves of U-Net and NAFNet based diffusion are illustrated in Figure 6. Note that the modified NAFNet significantly outperforms the U-Net backbone on the shadow removal task.

4.3. Improved Training Strategies

In this section, we discuss the main factors that affect the training process of diffusion-based restoration. The analysis

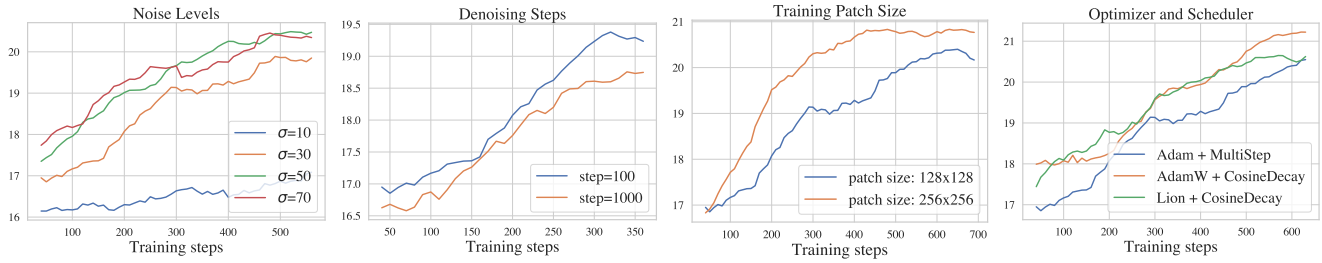


Figure 5. Validation PSNR curves of different training strategies on the real-world shadow removal task. All models use the same modified NAFNet backbone. We show that slightly changing these parameters can lead to significant performance improvements.

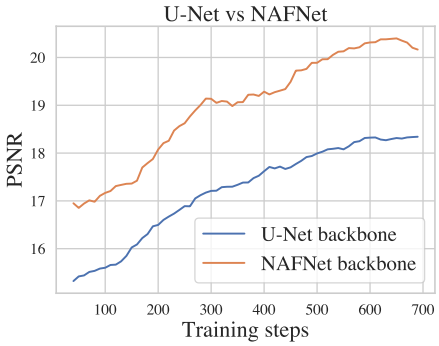


Figure 6. Comparison of learning curves between U-Net and the modified NAFNet backbone on the shadow removal dataset.

is performed using the real-world shadow removal task.

Noise levels. The noise level (i.e., the stationary variance λ from Section 3) can play an important role when it comes to the performance of diffusion models [41, 46]. In image restoration we often recover an HQ image directly from the LQ image rather than from pure noise, which means that a standard Gaussian for the terminal state is not necessary. As shown in Figure 5a, we compare four different noise levels $\sigma = \{10, 30, 50, 70\}$ on the shadow removal task. The training curves show that setting $\sigma = 50$ or $\sigma = 70$ is more stable than that with a small noise level, i.e. $\sigma = 10$.

Denoising steps. Several works propose to use long-step pretrained weights but generating images using fewer steps [5, 40, 56], which indeed improves the sample efficiency but however also at the cost of decreasing the image quality. In image restoration, we have to re-train diffusion models from scratch for all tasks. Since IR-SDE has a stable and robust learning process, we consider to directly adjust the denoising steps in training while maintaining the performance. Figure 5b compares the training curves of IR-SDE with 100 and 1000 denoising steps. We find that using fewer denoising steps can result in comparable—and sometimes even better—restoration performance.

Training patch sizes. A common practice is that training with large patches can improve the image restoration performance [36,37]. But none of the existing works discussed the effect of patch sizes in training diffusion models. Here we present the comparison of training IR-SDE with patch size 128×128 and 256×256, as shown in Figure 5c. As can be observed, training with large patches performs much better, which is consistent with other CNN/Transformer based image restoration approaches.

Optimizer/scheduler. A good optimizer with a proper learning rate scheduler is also important to the performance. As an example, simply adding a cosine decay scheduler can improve the accuracy by 0.5% for ResNet-50 on the ImageNet classification task [26]. To find out which optimizer better matches the diffusion model, we provide three comparisons including 1) Adam + multi-step decay, 2) AdamW [39] + cosine decay, 3) Lion [9] + cosine decay. The results in Figure 5d show that both AdamW and Lion perform slightly better then the Adam optimizer with multi-step learning rate decay.

5. Experiments

We evaluate Refusion on various image restoration tasks. In this section, we first briefly introduce several restoration tasks and their datasets, and then show the comparisons and results of our proposed method with other baselines. Our method achieves the best perceptual performance in the NTIRE 2023 Image Shadow Removal Challenge [62] and wins 2nd place in terms of overall performance.

5.1. Tasks and Datasets

Image Shadow Removal is the task of mapping shadow regions of an image to their shadow-free counterparts, which can enhance the image quality and benefit downstream computer vision tasks [45, 55, 79]. For the dataset, we follow the instructions in the NTIRE 2023 Shadow Removal Challenge [61,62] to use 1 000 pairs of shadow and shadow-free images for training and 100 shadow images for validation.

Stereo Image Super-Resolution is a problem stemming from

the growing popularity of dual cameras in modern mobile phones, and aims to recover high-quality images from paired low-quality left and right images with stereo correspondences [67, 71]. To train and evaluate our model, we use the dataset provided by the NTIRE 2023 Stereo Image Super-Resolution Challenge [64], which consists of 800 training stereo images and 112 validation stereo images from the Flickr1024 [66] dataset. All low-resolution images are generated by bicubic downsampling.

Bokeh Effect Transformation is an important task to computational photography, aiming to convert the image's Bokeh effect from the source lens to that of a target lens without harming the sharp foreground regions in the image [29, 30]. For this purpose, we consider the new dataset proposed in NTIRE 2023 Bokeh Effect Transformation challenge [15], in which 10 000 pairs of synthetic images with different lens information are used for training, and 500 images with source and target lens information are used for validation.

HR Non-Homogeneous Dehazing aims to perform defogging on extremely high-resolution images with heavy non-homogeneous fog, which is challenging to current dehazing approaches. Here we use a new dataset proposed in the NTIRE 2023 HR NonHomogeneous Dehazing competition [3]. This dataset has a large diversity of contents and is collected similar to NH-HAZE [1, 2], but with all images having $6000 \times 4000 \times 3$ pixels. In addition, it only contains 40 hazy/haze-free image pairs for training and 10 images for validation and testing. Since the clean images of the validation and test datasets are not accessible, we choose to only evaluate our method qualitatively on this task.

5.2. Implementation Details

For all experiments, we use the same setting as NAFNet. The batch sizes are set to 8 and the training patches are 256 $\times$ 256 pixels. We use the Lion optimizer [9] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 3×10^{-5} and decayed to 1e-7 by the Cosine scheduler. The noise level is fixed to 50 and the number of diffusion denoising steps is set to 100 for all tasks. We also augment the training data with random horizontal flips and 90 degree rotations. All models are implemented with PyTorch [48] and trained on a single A100 GPU for about 3 days.

For shadow removal and stereo super-resolution, we use the normal diffusion strategy and set the training iterations to 500 000. For HR dehazing and the Bokeh effect transformation, on the other hand, we incorporate the latent diffusion strategy. Specifically, we first train the U-Net on each dataset for 300 000 iterations, and then train the Refusion model based on the U-Net for 400 000 iterations. Here, all input patches are cropped to 1024×1024, while using U-Net to compress them to 128×128 pixels.

Table 1. Comparison of the proposed Refusion with IR-SDE [41] and NAFSSR [13] on the stereo super-resolution validation dataset. The proposed Refusion achieves significantly better performance than IR-SDE across all metrics, and outperforms NAFSSR in terms of perceptual scores.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	Runtime↓
NAFSSR [13]	23.81	0.7247	0.335	34.86	5.2s
IR-SDE [41]	20.34	0.5841	0.197	25.57	91.3s
Refusion	21.21	0.6336	**0.155**	**22.43**	64.1s

Table 2. Comparison of the proposed Refusion with IR-SDE [41], DHAN [16] and an L1 loss trained U-Net baseline on the shadow removal dataset. Our proposed Refusion achieves the best restoration performance overall.

Method	PSNR↑	SSIM↑	RMSE↓	LPIPS↓	FID↓	Runtime↓
DHAN [16]	20.42	0.6986	24.29	0.247	109.35	0.4s
IR-SDE [41]	20.30	0.6639	24.63	0.152	74.35	175.8s
U-Net baseline	20.69	0.7172	23.55	0.236	102.1	1.62s
Refusion	21.88	0.6977	20.53	**0.121**	**56.22**	38.4s

Table 3. Comparison of our methods with Restormer [72] on the Bokeh Effect Transformation dataset. Our Refusion with latent strategy can also achieve good performance.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	Runtime↓
Restormer [72]	41.12	0.9779	0.067	46.72	2.0s
Refusion	39.81	0.9615	0.053	**20.38**	36.0s
Latent Refusion	40.24	0.9721	**0.047**	24.25	6.9s

5.3. Experimental Results

Since our Refusion method is proposed for realistic image restoration, we use the Learned Perceptual Image Patch Similarity (LPIPS) [78] and Fréchet inception distance (FID) score [27] as the main evaluation metrics, but we also report PSNR and SSIM for reference. For shadow removal, we further report the RMSE metric as previous approaches [10, 16, 23]. Moreover, for each task, we also provide the runtime comparison to show the computational efficiency of our method against other baselines.

Stereo Image Super-Resolution. The quantitative comparison of our model with NAFSSR [13] and IR-SDE [41] is shown in Table 1. NAFSSR achieves the best PSNR and SSIM scores, but performs inferior to IR-SDE and our Refusion in terms of LPIPS and FID. The proposed Refusion significantly improves the performance of IR-SDE across all metrics, demonstrating the effectiveness of our improved training strategies. Our method runs slower than NAFSSR, which directly predicts HQ images from LQ images, but clearly improves the runtime of our main baseline IR-SDE. The visual results are illustrated in Figure 8. Our Refusion

Figure 7. Visual results of our method and the U-Net baseline on the shadow removal task. Top row shows the input images and second row shows the results produced by the U-Net baseline. Bottom row shows the shadow-free results generated by our method.

LR Image NAFSSR **Refusion**

Figure 8. Visual results of our method and NAFSSR [13] on the stereo super-resolution dataset.

produces sharper and clearer images than NAFSSR.

Image Shadow Removal. For this task, we compare our method with IR-SDE and a U-Net baseline model which uses the same network architecture as IR-SDE but is trained to directly predict HQ images via an L_1 loss. We also compare our method with DHAN [16], a well-established shadow removal model. The quantitative results are shown in Table 2. The proposed Refusion clearly achieves the best restoration performance overall. Moreover, Refusion runs significantly faster than IR-SDE. In the experiment, we find that all training image pairs have slight position shifts and that the luminance of each input and ground truth image is different, which may cause the L_1 loss trained model to learn the shift and luminance rather than shadow removal.

Table 4. Comparison of the parameters and flops. Flops are calculated on an image with size 256×256. Note that the additional U-Net for latent Refusion also has 6M parameters and 70G Flops, but it will be only run once for an image at test time.

Method	IR-SDE [41]	Refusion	Latent Refusion
#Params	135.3M	131.4 M	131.4M
Flops	119.1G	63.4G	4.0G

Thus its performance is lower than our Refusion even in terms of PSNR. The qualitative comparison in Figure 7 also demonstrates superior performance of Refusion.

Bokeh Effect Transformation. We apply our latent Refusion model to this task, with an image downscale factor set to 4. As shown in Table 3, both of our methods achieve better perceptual performance than Restormer [72]. The latent Refusion achieves a better LPIPS score but slightly worse FID than Refusion. It also runs much faster than Refusion, getting close to the runtime of Restormer. The visual comparison is shown in Figure 9. As one can see, our method produces sharper results than Restormer when transforming the bokeh effect from blurry to clear. We also provide a comparison of model complexities in Table 4. As can be observed, the latent strategy reduces computation flops about $15\times$ compared to the original Refusion model, which significantly improves the applicability.

HR Non-Homogeneous Dehazing. The visual results of dehazing are shown in Figure 10. As can be observed, most fog is successfully removed by our latent Refusion model. Note that all images in the HR dehazing dataset have $6000 \times 4000 \times 3$ pixels. With such large image sizes, IR-

Figure 9. Visual results of our method and Restormer [72] on the Bokeh effect transformation task. Top row shows the input images and second row shows the Restormer's results. Last row shows the transformed results generated by our method. In addition, lens transform information is shown in the bottom. 'S. f/1.8 → C. f/16' means the image is transformed from *Sony50mmf1.8BS* to *Canon50mmf16.0BS*.

Figure 10. Visual results of our method on the HR Non-Homogeneous Dehazing task. Top and bottom row are inputs and outputs.

SDE and Refusion are not even able to process a complete image at test time, and other diffusion models which use additional self-attention mechanisms would be even more computationally expensive. By performing the restoration in a low-resolution latent space, our latent Refusion model can be applied also in this highly challenging setting.

6. Conclusion

In this paper, we present several techniques to improve the applicability of diffusion-based image restoration. The resulting model, named *Refusion*, is successfully applied to various image restoration tasks and it achieves the best perceptual performance in the NTIRE 2023 Shadow Removal Challenge. To process large-size images, we further propose a U-Net based latent Refusion model that compresses the input image to a low-resolution latent space in which it performs diffusion to improve the model efficiency. Since input image information is also captured in the hidden vectors connected to the decoder, we are able to recover images with more accurate details. Our latent Refusion model can even run on images of size $6000 \times 4000 \times 3$ pixels.

Acknowledgements This research was partially supported by the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by Knut and Alice Wallenberg Foundation, by the project *Deep probabilistic regression – new models and learning algorithms* (contract number: 2021-04301) funded by the Swedish Research Council, and by the *Kjell & Märta Beijer Foundation*. The computations were enabled by the *Berzelius* resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 444–445, 2020. 6

[2] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2020 challenge on non-homogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 490–491, 2020. 6

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2023 challenge on non-homogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 6

[4] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022. 2

[5] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 5

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 2, 3, 4

[8] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 2

[9] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 5, 6

[10] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2021. 6

[11] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4896–4906, 2021. 2

[12] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 2

[13] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2022. 6, 7

[14] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022. 2

[15] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Tom E Bishop, and Radu Timofte. Lens-to-lens bokeh effect transformation. ntire 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 6

[16] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10680–10687, 2020. 6, 7

[17] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions. *arXiv preprint arXiv:2209.05442*, 2022. 2

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[20] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[21] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 2

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[23] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2021. 6

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[26] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567, 2019. 5

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 6

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 4

[29] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 418–419, 2020. 6

[30] Andrey Ignatov, Radu Timofte, Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, Jian Cheng, Juewen Peng, et al. Aim 2020 challenge on rendering realistic bokeh. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 213–228. Springer, 2020. 6

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[32] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 2

[33] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021. 2

[34] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[35] Wenyi Lian and Shanglian Peng. Kernel-aware burst blind super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4892–4902, 2023. 2

[36] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2, 5

[37] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. Revisiting rcan: Improved training for image super-resolution. *arXiv preprint arXiv:2201.11279*, 2022. 5

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 5

[41] 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. 2, 3, 5, 6, 7

[42] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17652, 2022. 2

[43] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2022. 2

[44] Ziwei Luo, Youwei Li, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, and Shuaicheng Liu. Fast nearest convolution for real-time efficient image super-resolution. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 561–572. Springer, 2023. 2

[45] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 26(8):1079–1087, 2004. 5

[46] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 5

[47] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 4

[50] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 2

[51] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 154–169. Springer, 2016. 2

[52] 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. 4

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[54] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[55] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In *2010 20th International Conference on Pattern Recognition*, pages 141–144. IEEE, 2010. 5

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3, 4

[58] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 2

[59] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 3

[60] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 2

[61] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5

[62] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. Ntire 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[64] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2023. 6

[65] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[66] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[67] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–775, 2021. 6

[68] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 2, 3

[69] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 2

[70] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[71] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020. 6

[72] 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*, pages 5728–5739, 2022. 2, 3, 6, 7, 8

[73] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 2

[74] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 2

[75] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021. 2

[76] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2

[77] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 2

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[79] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):611–624, 2018. 5

[80] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2

[81] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 2