Masked Image Training for Generalizable Deep Image Denoising

Haoyu Chen[1*], Jinjin Gu[2,3*], Yihao Liu[2,4,5], Salma Abdel Magid[7],
Chao Dong[2,4], Qiong Wang[6], Hanspeter Pfister[7], Lei Zhu[1,8†]

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Shanghai AI Lab [3]The University of Sydney
[4]ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[5]University of Chinese Academy of Sciences [6]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[7]Harvard University [8]The Hong Kong University of Science and Technology

Project page: https://github.com/haoyuc/MaskedDenoising

Abstract

When capturing and storing images, devices inevitably introduce noise. Reducing this noise is a critical task called image denoising. Deep learning has become the de facto method for image denoising, especially with the emergence of Transformer-based models that have achieved notable state-of-the-art results on various image tasks. However, deep learning-based methods often suffer from a lack of generalization ability. For example, deep models trained on Gaussian noise may perform poorly when tested on other noise distributions. To address this issue, we present a novel approach to enhance the generalization performance of denoising networks, known as masked training. Our method involves masking random pixels of the input image and reconstructing the missing information during training. We also mask out the features in the self-attention layers to avoid the impact of training-testing inconsistency. Our approach exhibits better generalization ability than other deep learning models and is directly applicable to real-world scenarios. Additionally, our interpretability analysis demonstrates the superiority of our method.

1. Introduction

Image denoising is a crucial research area that aims to recover clean images from noisy observations. Due to the rapid advancements in deep learning, many promising image denoising networks have been developed. These networks are typically trained using images synthesized from a pre-defined noise distribution and can achieve remarkable performance in removing the corresponding noise. However, a significant challenge in applying these deep models to real-world scenarios is their generalization ability. Since the real-world noise distribution can differ from that observed during training, these models often struggle to generalize to such scenarios.

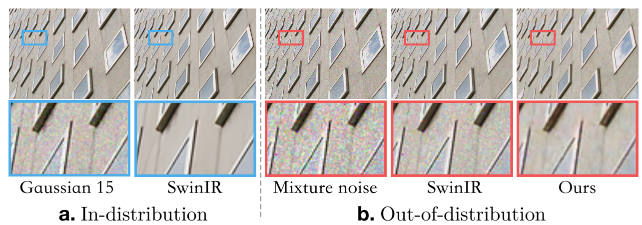


Figure 1. We illustrate the generalization problem of denoising networks. We train a SwinIR model on Gaussian noise with $\sigma = 15$. When tested on the same noise, SwinIR demonstrates outstanding performance. However, when applied to out-of-distribution noise, *e.g.*, the mixture of various noise. SwinIR suffers from a huge performance drop. The model trained by the proposed *masked training* method maintains a reasonable denoising effect, despite aldo being trained on Gaussian noise.

More specifically, most existing denoising works train and evaluate models on images corrupted with Gaussian noise, limiting their performance to a single noise distribution. When these models are applied to remove noise drawn from other distributions, their performance drastically drops. Figure 1 shows an example. The research community has become increasingly aware of this generalization issue of deep models in recent years. As a countermeasure, some methods [80] assume that the noise level of a particular noise type is unknown, while others [5, 68] attempt to improve the performance in real-world scenarios by synthesizing or collecting training data closer to the target noise or directly performing unsupervised training on the target noise [11, 71]. However, none of these methods substantially improve the generalization performance of denoising networks, and they still struggle when the noise distribution is mismatched [1]. The generalization issue of

*Haoyu Chen and Jinjin Gu contribute equally to this work.
†Lei Zhu (leizhu@ust.hk) is the corresponding author.

deep denoising still poses challenges to making these methods broadly applicable.

In this work, we focus on improving the generalization ability of deep denoising models. We define generalization ability as the model's performance on noise different from what it observed during training. We argue that the generalization issue of deep denoising is due to the overfitting of training noise. The existing training strategy directly optimizes the similarity between the denoised image and the ground truth. The intention behind this is that the network should learn to reconstruct the texture and semantics of natural images correctly. However, what is often overlooked is that the network can also reduce the loss simply by overfitting the noise pattern, which is easier than learning the image content. This is at the heart of the generalization problem. Even many popular deep learning methods exacerbate this overfitting problem. When it comes to noise different from that observed during training, the network exhibits this same behavior, resulting in poor performance.

In light of the preceding discussion, our study seeks to improve the generalization performance of deep denoising networks by directing them to learn image content reconstruction instead of overfitting to training noise. Drawing inspiration from recent masked modeling methods [4, 20, 34, 69], we employ a masked training strategy to explicitly learn representations for image content reconstruction, as opposed to training noise. Leveraging the properties of image processing Transformers [15,45,78], we introduce two masking mechanisms: the *input mask* and the *attention mask*. During training, the input mask removes input image pixels randomly, and the network reconstructs the removed pixels. The attention mask is implemented in each self-attention layer of the Transformer, enabling it to learn the completion of masked features dynamically and mitigate the distribution shift between training and testing in masked learning. Although we use Gaussian noise for training – similar to previous works – our method demonstrates significant performance improvements on various noise types, such as speckle noise, Poisson noise, salt and pepper noise, spatially correlated Gaussian noise, Monte Carlo-rendered image noise, ISP noise, and complex mixtures of multiple noise sources. Existing methods and models have yet to effectively and accurately remove all these diverse noise patterns.

2. Related Works

Image Denoising approaches very broadly lie in two categories: traditional model-based and data-driven deep-learning-based. Traditional methods are usually based on modeling image priors to recover image content contaminated by noise [7, 19, 23, 32, 53]. These methods usually do not impose too many constraints on the type of noise, and have been proven to be applicable to a variety of noise, with good generalization performance [1]. However, these methods are not satisfactory for the reconstruction of image content. In recent years, the paradigm of denoising has gradually shifted to data-driven methods based on deep learning methods [13]. Many techniques have been proposed to improve the capabilities of the denoising networks continuously, *e.g.*, residual networks [39,80,81], dense networks [37, 86], recursive networks [9, 48, 63], multi-scale [21, 31, 76], encoder-decoder [16, 54, 73], attention operations [84, 85], self-similarity [35], and non-local operations [43, 44, 58]. Since 2020, the paradigm of vision network design has gradually shifted from CNNs to Transformers [22]. Vision Transformers treat input pixels as tokens and use self-attention operations to process interactions between these tokens. Inspired by the success of vision Transformers, many attempts have been made to employ Transformers for low-level vision tasks [10,14,15,45,62,67,70,74,77,78] During the development of these models, the noise pattern used for training is often consistent with the testing one. The factor that determines its denoising performance is the fitting ability of the network, in other words, the ability of the network to overfit to the training noise. However, a better network does not mean a better generalization ability of the denoising model. As we will show in the experiment section, a more efficient network even indicates worse generalization performance.

Generalization Problem in low-level vision often arises when the testing degradation does not match the training degradation, *e.g.*, different downsampling kernel in super-resolution [30, 40, 47]. We typically develop deep denoising models based on Gaussian noise in the laboratory setting. However, noise in the real-world is mostly non-Gaussian. Models trained on Gaussian noise fail in these non-Gaussian scenarios. There are two main categories of solutions to this problem. The first is to make training datasets with noise modeling as close to reality as possible during development, *e.g.*, synthesizing real noise according to physical system modeling [5, 68], learning to generate real noise [11, 24, 71], collecting real noise – clean image pairs for training [1, 33, 41, 57]. Although the models obtained by these methods can improve the effect on the target noise, they still cannot generalize to out-of-distribution noise. Another category of solutions is to develop "blind" denoising models, which are supposed to deal with unknown noise [41, 72, 80]. These methods usually simply assume that the noise level is unknown, or train on a large amount of noise types [79], which also fails to generalize to other noise not present in the training set. Few workd have been proposed to study the reasons for the lack of generalization ability in low-level vision [40]. Liu *et al*. [49] argue that networks tend to overfit to degradations and show degradation "semantics" inside the network. The presence of these representations often means a decrease in general-

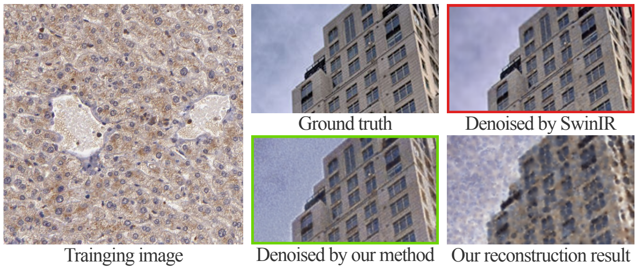


Figure 2. SwinIR, when trained solely on immunohistochemistry images with Gaussian noise, can still denoise natural images. This observation supports the assertion that most existing methods perform denoising primarily through overfitting the training noise. In contrast, our approach emphasizes reconstructing natural image textures and edges observed in the training set on natural images, rather than relying on noise overfitting for denoising. This distinction underlines the fundamental difference between our method and previous approaches. "Our reconstruction result" refers to using our model but taking masked images as input.

ization ability. The utilization of this knowledge can guide us to analyze and evaluate the generalization performance [50]. Apart from that, few works have been proposed to improve the generalization ability of denoising models.

Masked modeling for language [6, 20, 59, 60] is successful for learning pre-trained representations that generalize well to various downstream tasks. These methods mask out a portion of the input sequence and train models to predict the missing content. A similar approach can also be applied to the vision model pre-training. Masked image models learn representations from corrupted images. The earliest attempts in this regard can be traced back at least to the denoising auto-encoder [66]. Since then, many works have used predicting missing parts of images to learn efficient image representations [4, 12, 34, 56, 69]. However, there have been few successful attempts to apply masked image modeling to low-level vision, even though the masked pretraining method is in the form of low-level vision tasks.

3. Method

Our objective is to create denoising models capable of generalizing to noise not encountered in the training set. In this section, we first discuss our motivation before delving into the specifics of our masked training method.

Motivation. When training a deep network on a large number of images, the expectation is for the network to learn to discern the rich semantics of natural images from noise-contaminated test cases. However, several studies have noted that the semantics and knowledge acquired by low-level vision networks differ significantly from our expectations [29, 49, 50, 52]. We argue that the poor generalization ability of denoising models results from our training method, which leads the model to *focus on overfitting the training noise rather than learning image reconstruction*. We conduct a simple experiment for verification. We trained a SwinIR denoising network [45] using images that greatly differ from natural images (immunohistochemistry images [65]). We synthesized training data pairs using Gaussian noise, and then assessed the model's performance on *natural images* with Gaussian noise. According to our hypothesis, if the model learns the content and reconstruction of image semantics from the training set, it should not perform well on natural images, as it has not been exposed to any. If the model is simply overfitting the noise, the model can remove the noise even if the images are different, as the model mainly relies on detecting the noise for denoising.

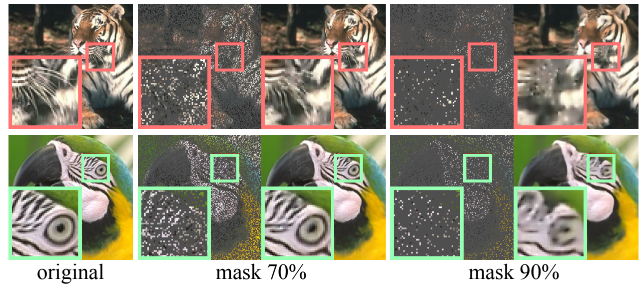


Figure 3. The illustration of the proposed mask-and-complete training strategy. Even if a large number of pixels are masked, the model can still reconstruct the input to some extent.

The results are presented in Figure 2. As observed, the SwinIR trained on immunohistochemistry images can still denoise and reproduce the natural image. This supports our conjecture regarding generalization ability, indicating that most existing methods perform denoising by overfitting the training noise. Consequently, when the noise deviates from the training conditions, the denoising performance of these models declines significantly.

This observation also inspires our approach to developing deep denoising models with improved generalization ability. We aim for the model to learn the reconstruction of image textures and structures, rather than focusing only on noise. In this paper, we propose a new masked training strategy for denoising networks. During training, we mask out a portion of the input pixels and then train the deep network to complete them, as shown in Figure 3. Our approach emphasizes reconstructing natural image textures and edges observed in the image, rather than overfitting noise. In Figure 2 we also show the results of our method. It is evident that our approach seeks to reconstruct the immunohistochemistry image texture from the training set on the testing natural image, instead of relying on noise overfitting for denoising. This demonstrates the potential of this idea in improving generalization performance. By training our method on natural images, it will concentrate on reconstructing the content of natural images, aligning with our

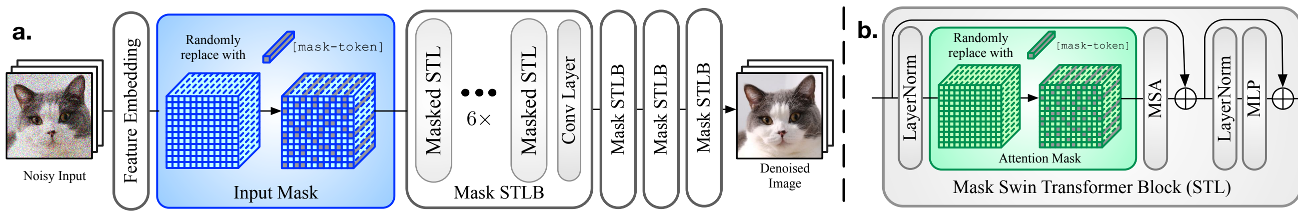


Figure 4. The transformer architecture of our proposed masked image training. We make a minimal change to the original SwinIR architecture – the input mask operation and the attention masks. Other micro-designs are not essentially different from other Transformers.

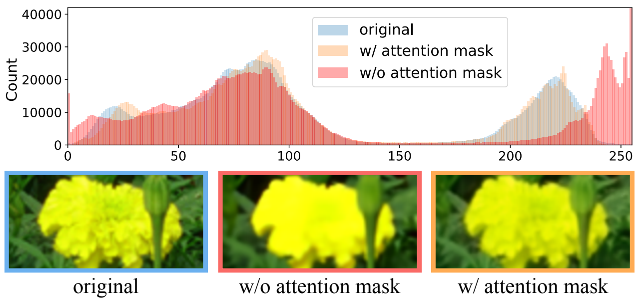


Figure 5. Quantitative effect of the attention mask. The histogram differences are also shown above.

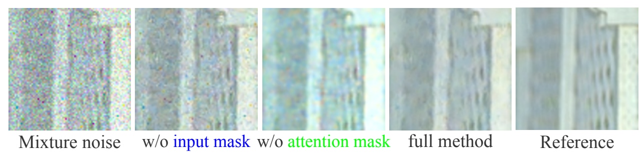


Figure 6. The effectiveness of the input mask and attention mask. Note that the brightness of the image is wrong *w/o* attention mask.

Input Mask	Attention Mask	PSNR	SSIM
	✓	29.17	0.8227
✓		26.96	0.8202
✓	✓	**29.74**	**0.8672**

Table 1. The importance of using different mask operations.

Mix. noise on CBSD68 [55]		
Ratio (%)	PSNR	SSIM
65	29.57	0.8657
75	**29.76**	**0.8678**
85	28.84	0.8548

Table 2. Ablation on the attention mask ratio.

core concept of employing deep learning for low-level vision tasks.

The Transformer Architecture. Our approach exploits the excellent properties of visual Transformers, so we first describe the basic Transformer backbone used in this study. The shifted window mechanism is proven to be flexible and effective for image processing tasks [15, 45, 78]. We only make minimal changes when applying it to the proposed masked training method without the loss of generality. This model is illustrated in Figure 4. Transformers divide the input signal into tokens and process spatial information using self-attention layers. In our method, a convolution layer with kernel size 1 is used as the feature embedding module to project the 3-channel pixel values into C-dimensional feature tokens. The 1×1 convolution layer ensures that pixels do not affect each other during feature embedding, which facilitates subsequent masking operations. These feature tokens are gathered with shape $H \times W \times C$, where H, W and C are the height, width and feature dimension. The shifted window mechanism first reshapes the feature maps of each frame to $\frac{HW}{M^2} \times M^2 \times C$ features by partitioning the input into non-overlapping $M \times M$ local windows, where $\frac{HW}{M^2}$ is the total number of windows. We calculate self-attention on the feature tokens within the same window. Therefore, M^2 tokens are involved in each standard self-attention operation, and we produce the local window feature $X \in \mathbb{R}^{M^2 \times C}$. In each self-attention layer, the query Q, key K and value V are calculated as $Q = XW^Q$, $K = XW^K$, $V = XW^V$, where $W^Q, W^K, W^V \in \mathbb{R}^{C \times D}$ are weight matrices, and D is the dimension of projected vectors. Then, we use Q to query K to generate the attention map $A = \texttt{softmax}(QK^T/\sqrt{D} + B) \in \mathbb{R}^{M^2 \times M^2}$, where B is the learnable relative positional encoding. This attention map A is then used for the weighted sum of M^2 vectors in V. The multi-head settings are aligned with SwinIR [45] and ViT [22].

Masked Training. Our masked training mainly consists of two aspects, the input mask and the attention mask. Although both are mask operations, the purpose of these two masks is different. We describe them separately.

The Input Mask randomly masks out the feature tokens embedded by the first convolution layer, and encourages the network to complete the masked information during training. The input mask explicitly constructs a very challenging inpainting problem, as shown in Figure 3. It can be seen that even if up to 90% of the pixel information is destroyed, the network can still reconstruct the target image to a certain extent. The method is very simple. Given the feature token tensor $\mathbf{f} \in \mathbb{R}^{H \times W \times C}$, we randomly replace the token with a `[mask token]` $\in \mathbb{R}^C$ with a probability p_{IM}, where p_{IM} is called the input mask ratio. The network is trained under the supervision of the l_1-norm of the reconstructed image and the ground truth. The `[mask token]` can be learnable and initialized with a $\mathbf{0}$ vector. But we actually found that the $\mathbf{0}$ vector itself is already a suitable choice. The existence of the input mask forces the network to learn to recognize and reconstruct the content of the image from very limited information.

The Attention Mask. We cannot build usable image pro-

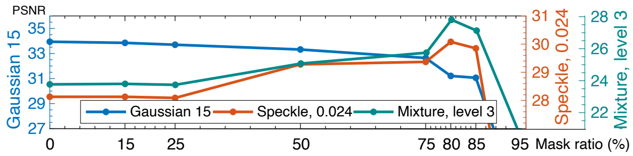


Figure 7. The trade-off of choosing different mask ratios. The performance drop on training noise is not significant until 75% masking ratio. Our performance gain on the noise outside the training set is greater than the performance loss on the training set.

cessing networks relying solely on the input mask operation. Because during testing, we will input uncorrupted images to retain enough information. At this time, due to the inconsistency between training and testing, the network will tend to increase the brightness of the output image. Such as the example in Figure 5. Since Transformer uses the self-attention operation to process spatial information, we can narrow the gap between training and testing by performing the same mask operation during the self-attention process. The specific mask operation is similar to the input mask, but a different attention mask ratio p_{AM} and `[mask token]` are used. When some tokens in the self-attention are masked, the attention operation will adjust to the fact that the information of these tokens is no longer reliable. Self-attention will focus on unmasked tokens in each layer and complete the masked information. This operation is difficult to implement on convolutional networks. Figure 5 shows the effect of the attention mask. As can be seen, the attention mask successfully makes the masked trained network work on the unmasked input image.

4. Experiments

Training Settings. For synthesizing training data, we sample the clean images from DIV2K [64], Flickr2K [46], BSD500 [3], and WED [51] during training. In our work, all the networks are trained using Gaussian noise with standard deviation $\sigma = 15$. Each input image is randomly cropped to a spatial resolution of 64×64, and the number of the total training iteration is 200K. We adopt the Adam optimizer [38] with β_1=0.9 and β_2=0.99 to minimize the L_1 pixel loss. The initial learning rate is set as 1×10^{-4} and reduced by half at the milestone of 100K iterations and 150K iterations. The batch size is set to 64.

Testing Noise. Since the training process utilizes the Gaussian noise, we evaluate the generalization performance of the models on six other synthetic noise: (1) Speckle noise, a type of noise that occurs during the acquisition of medical images or tomography images. (2) Poisson noise, a type of signal-dependent noise that occurs during the acquisition of digital images.(3) Spatially-correlated noise. This is to synthesize the complex artifact after denoising using a flawed algorithm. It is produced by filtering Gaussian noise with a 3×3 average kernel. Different standard deviations of the Gaussian noise indicate different noise levels. (4) Salt & pepper noise. (5) Image signal processing (ISP) noise. [5] proposes a method to synthesize realistic ISP noise during digital imaging. (6) Mixture noise obtained by mixing the above different types of noise with different levels [79]. The clean images are sampled from the benchmark datasets, including CBSD68 [55], Kodak24 [26], McMaster [82], and Urban100 [36]. We also include two real noise types in this work: the Smartphone Image Denoising Dataset (SIDD) [1] and Monte Carlo (MC) rendered image noise. For evaluation, we follow [27,28] and use the metrics PSNR, SSIM [51], and LPIPS [83] to evaluate the results. Since PSNR and SSIM are questioned in assessing the perceptual quality of images [27,28], we also use the LPIPS as an additional metric.

4.1. Resutls

Ablation Study. Table 1 and Figure 6 show the effectiveness of using different mask operations. As we can see, without the input mask, the model will lose its generalization ability, and cannot effectively remove the noise outside the training set. Without the attention mask, due to the training-testing inconsistency, the quantitative performance degrades significantly, and the output image will have the wrong brightness. In addition, even without the attention mask, the generalization ability of the model is not significantly affected, and most of the noise is still effectively removed. The input mask is the crucial factor in improving the model the generalization ability.

Table 3a shows the impact of the different input mask ratios. We test fixed ratios and random ratios from a uniform distribution. From our experiments, fixed ratios are less stable for training than randomly chosen from a range, and the performance is also worse. The best quantitative performance is achieved with random sampling ratios between 75% $\sim$ 85%. This is a trade-off between denoising generalization ability and the preservation of image details. As shown in Figure 7, smaller ratios are not enough for the network to learn the distribution of images because more noise patterns are preserved. The larger ratio improves the model generalization, as the model focuses more on reconstruction. But at the same time, some image details may be lost. For attention mask ratio, we show the effects in Table 2. The optimal ratios are around 75%.

The Generalization Performance. We evaluate our deep denoising method on synthetic noise, where our training noise follows a Gaussian distribution with a single noise level, but we test on multiple types of non-Gaussian noise to assess the model's generalization performance. In Figure 11, we compare our method with other state-of-the-art models based on their PSNR and SSIM scores. The results show that our model outperforms all the other models in terms of generalization performance. Particularly, as

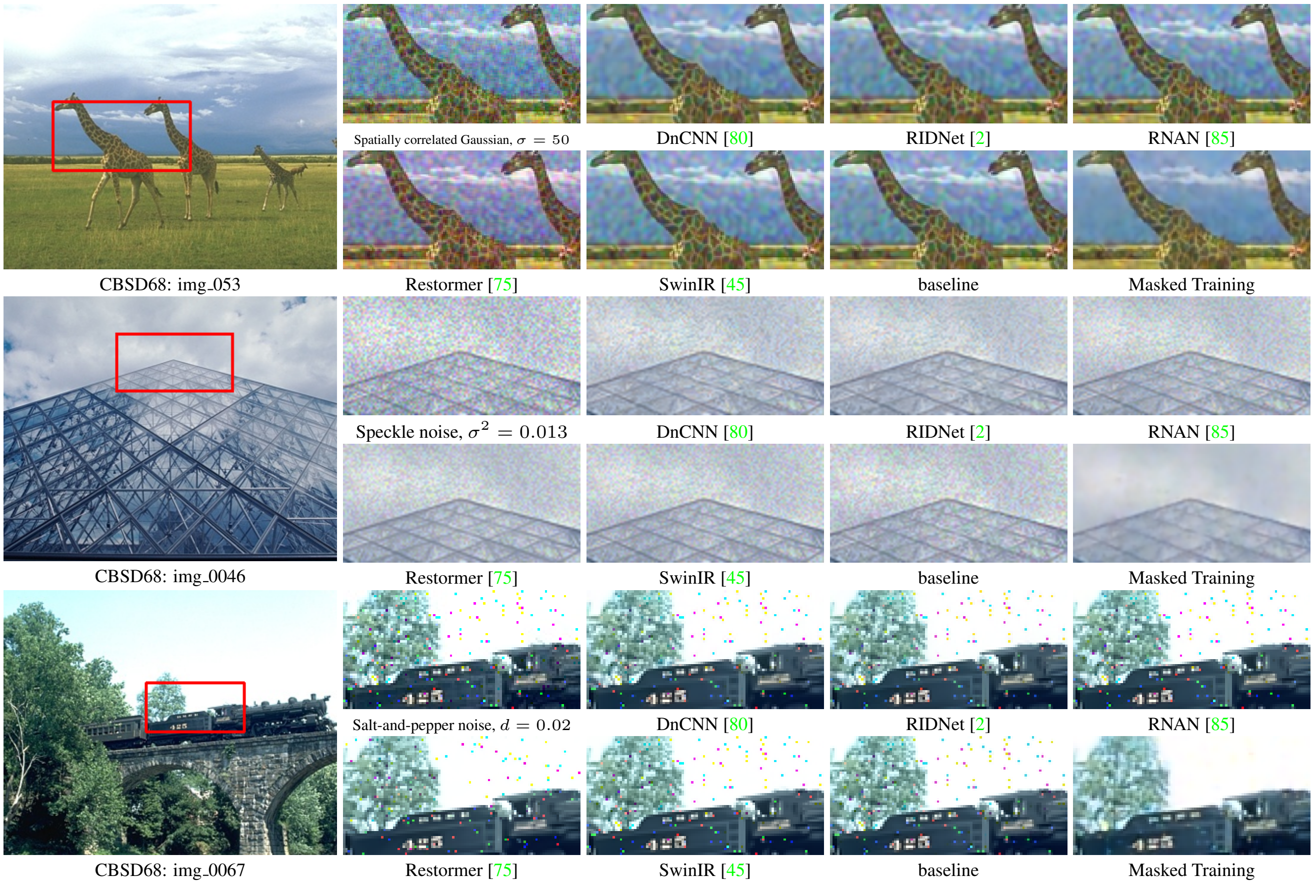


Figure 8. Visual comparison on out-of-distribution noise. When all other methods fail completely, our method is still able to denoise effectively. Please refer to the supplementary material to see more visual results.

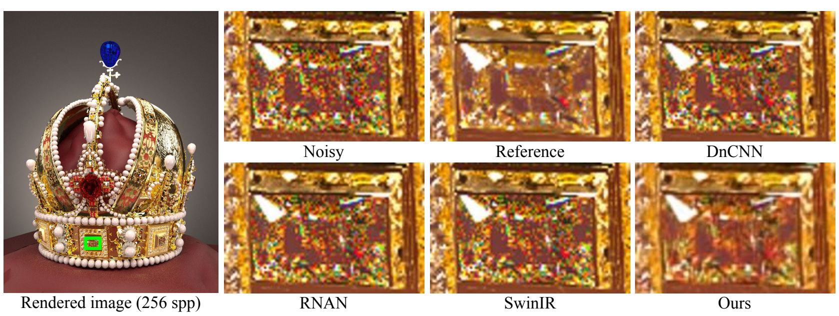


Figure 9. Visual results of denoising a Monte Carlo rendered image.

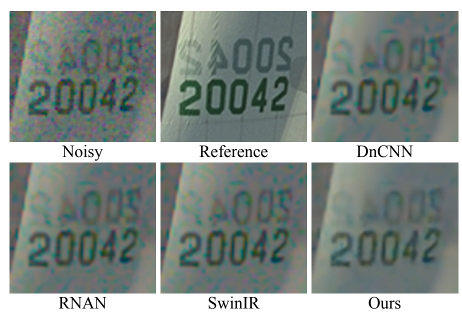


Figure 10. Results of ISP noise removal.

the noise level increases, our model exhibits a slower performance degradation and thus demonstrates better generalization. In contrast, other models suffer from significant performance drops when dealing with more severe noise. We also provide visual comparisons in Figure 8, where our model achieves remarkable denoising results even though it is trained only on Gaussian noise with a fixed standard deviation. In contrast, existing models tend to overfit the training noise and fail when facing unseen noise. More quantitative and qualitative results can be found in the supplementary material.

Evaluation on ISP noise. The removal of the ISP noise is of great application value. Brooks *et al.* [5] present a systematic approach for generating realistic raw data with ISP noise that can facilitate our research. We use the default parameter settings of the method proposed in [5] to synthesize ISP noise on the Kodak24 [26] dataset for testing. The results are shown in Figure 10 and Table 3c. Our method achieves superior results compared to all other methods. Notably, our method achieves a significant lead in LPIPS, indicating that our results exhibit better perceptual quality. Although DnCNN and our method obtain the same PSNR, our method still outperforms DnCNN in terms of SSIM and

Mix. noise on CBSD68 [55]		
Ratio (%)	PSNR	SSIM
75	29.17	0.8132
85	29.44	0.8545
95	19.60	0.7273
70–80	29.86	0.8593
75–85	**30.04**	**0.8756**
75–90	29.87	0.8728
75–95	29.26	0.8607
80–90	29.74	0.8672

(a) Abl. of input mask ratios.

	128 samples per pixel			64 samples per pixel		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DnCNN [80]	29.94	0.7883	0.2671	26.28	0.6779	0.4216
RIDNet [2]	29.96	0.7921	0.2548	26.27	0.6788	0.4122
RNAN [85]	29.86	0.7825	0.2702	26.26	0.6743	0.4290
SwinIR [45]	29.32	0.7627	0.2943	26.14	0.6651	0.4485
Restormer [75]	24.98	0.6598	0.4575	24.59	0.5880	0.5375
Dropout [40]	28.85	0.7753	0.2941	26.10	0.6696	0.4443
baseline	29.68	0.7738	0.2851	25.91	0.6535	0.4564
Ours	**30.62**	**0.8500**	**0.2254**	**28.25**	**0.7694**	**0.3348**

(b) Quantitative comparison on Monte Carlo rendered image denoising.

	Synthetic ISP noise [5]		
Method	PSNR	SSIM	LPIPS
DnCNN [80]	**29.44**	0.7857	0.3083
RIDNet [2]	28.75	0.7446	0.3696
RNAN [85]	28.47	0.7243	0.3601
SwinIR [45]	28.39	0.7079	0.3346
Restormer [75]	19.31	0.4982	0.6556
Dropout [40]	28.39	0.7816	0.2621
baseline	28.89	0.7595	0.2917
Ours	**29.44**	**0.7920**	**0.2368**

(c) Comparison on synthetic ISP noise.

Table 3. We train all the models on Gaussian noise, $\sigma = 15$. All the testing noise is out of the training set, therefore the results can show the models' generalization performance on different unseen noise.

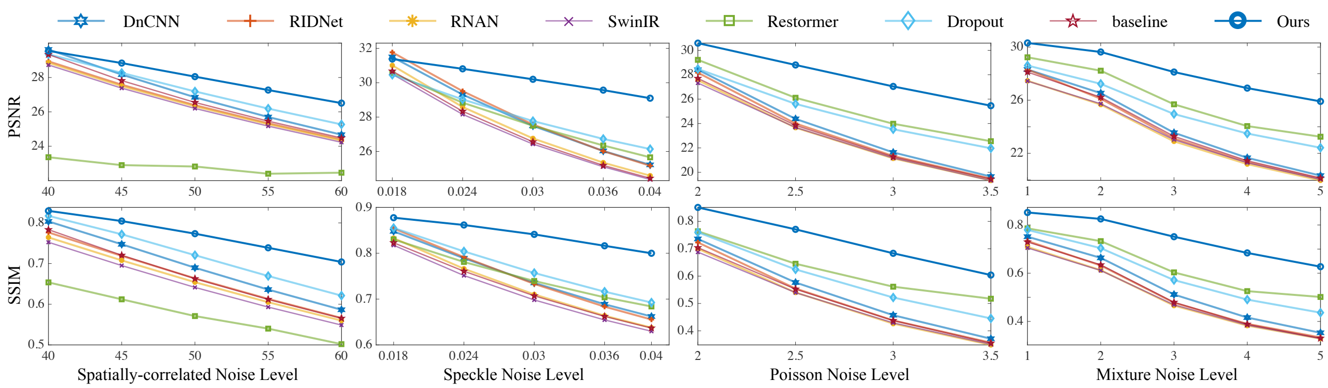


Figure 11. Performance comparisons on four noise types with different levels on the Kodak24 dataset [26]. All models are trained only on Gaussian noise. Our masked training approach demonstrates good generalization performance across different noise types. We involve multiple types and levels of noise in testing, the results cannot be shown here. More results are shown in the supplementary material.

LPIPS. Furthermore, as evident from Figure 10, DnCNN's results still contain visible noise, while our method effectively removes the noise.

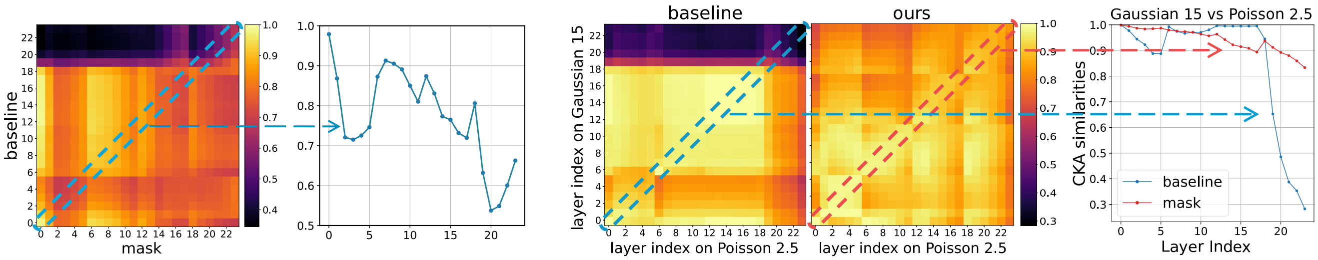
Evaluation on Monte Carlo rendering noise. Monte Carlo denoising is a vital component of the rendering process since the widespread use in the industry of Monte Carlo rendering algorithms [8, 17, 42]. We use the test dataset proposed by [25] for Monte Carlo rendered image denoising. The test images were rendered in 128 samples-per-pixel (spp) and 64 spp. The lower the spp, the more severe the noise of the image. In order to adapt the test set to our model, we first convert the data set to sRGB color space by tone mapping. Figure 9 and Table 3b show the denoising results. Our method outperforms all methods on both 128spp and 64spp settings. In Figure 9, the existing methods fail completely because of poor generalization. Our model is still able to remove this noise, demonstrating the wide applicability of our method.

4.2. Generalization Analysis

Training curve. Figure 13 shows the training curves of the model with and without the proposed masked training. The models are trained using only Gaussian noise. The baseline method has a significant overfitting problem. The performance of our method gradually improves with training without overfitting.

CKA analysis. To investigate how masked training differs from normal training strategy, we utilize the centered kernel alignment (CKA) [18, 61] to analyze the differences between network representations obtained from those two training methods. Due to the limited space, we describe the detail of CKA in supplementary. In Figure 12, we present our key findings. Specifically, Figure 12 (a) shows the cross-model comparison between the baseline model and our masked training model. We observe a significant difference between the two models in terms of their feature correlations in the deeper layers. Specifically, the features of the deeper layers of the baseline model exhibit low correlations with all layers of our model. This finding suggests that these two training methods exhibit inconsistent learning patterns for features, especially for the deeper layers.

To explore how the models perform on different noise types, Figure 12 (b) shows the cross-noise comparison between in-distribution noise and out-of-distribution noise, such as Gaussian and Poisson noise. For the baseline model, we observe a low correlation between different noise types in the deep layers, indicating that the network processes these two types of noise in different ways for the deep layers. This trend holds for other types of noise as well. This phenomenon may be due to the baseline approach causing



a. cross-model comparison on Gaussian 15

b. cross-model comparison between different noise

Figure 12. CKA similarity to analyze the representation similarity of network layers.

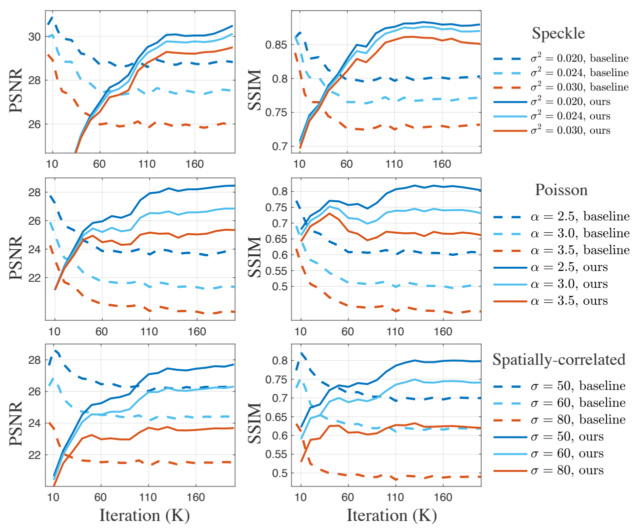


Figure 13. The testing curves on different noise types and levels.

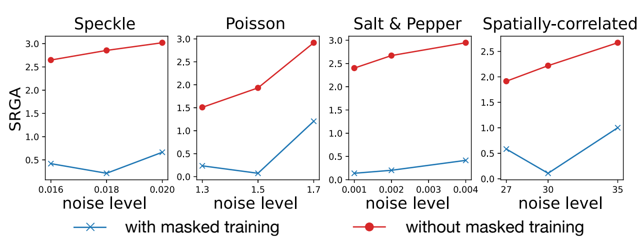


Figure 14. Comparing generalization ability with the SRGA metric. A lower SRGA value indicates better generalization ability.

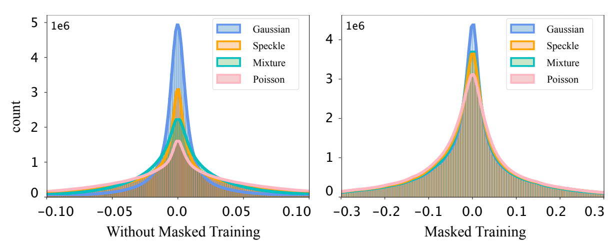


Figure 15. The distribution of baseline model features is biased across different noise types. Our method produces similar feature distributions across different noise.

the deep layers of the model to overfit to the patterns of the training set, thereby limiting their generalization capabilities to handle different noise types. In contrast, the high correlation between adjacent layers in our masked training model suggests that the model's representation of two different noise types is similar. Our masked training forces the network to learn the underlying distribution of the images themselves, which makes the model more robust to different noise types and enhances its generalization capability.

Quantification of generalization performance. Liu *et al*. [49, 50] suggest that model generalization ability can be measured by measuring the consistency of the model's representations across different types of noise. They also propose a generalization assessment index for low-level vision networks called SRGA [50]. It is a non-parametric and non-learning metric which exploits the statistical characteristics of internal features of deep networks. The lower the value of SRGA, the better the generalization ability. In our case, we use Gaussian noise as the reference and other types of noise for testing. Figure 14 shows the SRGA results. Inspired by [50], we visualize the distributions of deep features on different noise types, shown in Figure 15. We can see that for the baseline model, the feature distributions under different noise types deviate from each other significantly. For the model *w/* masked training, the deep feature distributions of different noise types are close to each other. This confirms the effectiveness of our method.

5. Conclusion and Limitations

This work presents a masked training method to improve the generalization performance of deep learning-based image denoising models. The limitation of our method is that the mask operation inevitably loses information. How to preserve more details needs to be explored in future work. Our approach is a step towards developing more robust models for real-world applications.

Acknowledgment. This work is supported in part by Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), the National Natural Science Foundation of China under Grant (62276251), NSFC General Project 62072452, Guangdong Provincial Basic and Applied Basic Research Fund- Regional Joint Fund (Project reference number 2020B1515130004), the Joint Lab of CAS-HK, and the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2020356).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 1, 2, 5

[2] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3155–3164, 2019. 6, 7

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2010. 5

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 3

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 1, 2, 5, 6, 7

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[7] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005. 2

[8] Brent Burley, David Adler, Matt Jen-Yuan Chiang, Hank Driskill, Ralf Habel, Patrick Kelly, Peter Kutz, Yining Karl Li, and Daniel Teece. The design and evolution of disney's hyperion renderer. *ACM Transactions on Graphics (TOG)*, 37(3):1–22, 2018. 7

[9] Chang Chen, Zhiwei Xiong, Xinmei Tian, and Feng Wu. Deep boosting for image denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018. 2

[10] 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 *CVPR*, 2021. 2

[11] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3155–3164, 2018. 1, 2

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 3

[13] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016. 2

[14] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. *arXiv preprint arXiv:2303.06373*, 2023. 2

[15] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NIPS*, 2022. 2, 4

[16] 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

[17] Per Christensen, Julian Fong, Jonathan Shade, Wayne Wooten, Brenden Schubert, Andrew Kensler, Stephen Friedman, Charlie Kilpatrick, Cliff Ramshaw, Marc Bannister, et al. Renderman: An advanced path-tracing architecture for movie rendering. *ACM Transactions on Graphics (TOG)*, 37(3):1–21, 2018. 7

[18] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012. 7

[19] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[21] Nithish Divakar and R Venkatesh Babu. Image denoising via cnns: An adversarial approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 80–87, 2017. 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, 4

[23] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006. 2

[24] Ruicheng Feng, Jinjin Gu, Yu Qiao, and Chao Dong. Suppressing model overfitting for image super-resolution networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[25] Arthur Firmino, Jeppe Revall Frisvad, and Henrik Wann Jensen. Progressive denoising of monte carlo rendered images. In *Computer Graphics Forum*, volume 41, pages 1–11. Wiley Online Library, 2022. 7

[26] Rich Franzen. Kodak lossless true color image suite. source: http://r0k.us/graphics/kodak/, 1999. 5, 6, 7

[27] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for percep-

tual image restoration: A new dataset, benchmark and metric. *arXiv preprint arXiv:2011.15002*, 2020. 5

[28] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 5

[29] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 3

[30] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 2

[31] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2511–2520, 2019. 2

[32] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014. 2

[33] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019. 2

[34] 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, 3

[35] Xiaowan Hu, Ruijun Ma, Zhihong Liu, Yuanhao Cai, Xiaole Zhao, Yulun Zhang, and Haoqian Wang. Pseudo 3d autocorrelation network for real image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16175–16184, 2021. 2

[36] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 5

[37] Xixi Jia, Sanyang Liu, Xiangchu Feng, and Lei Zhang. Focnet: A fractional optimal control network for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6054–6063, 2019. 2

[38] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[39] Filippos Kokkinos and Stamatios Lefkimmiatis. Deep image demosaicking using a cascade of convolutional residual denoising networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–319, 2018. 2

[40] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6002–6012, 2022. 2, 7

[41] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019. 2

[42] Christopher Kulla, Alejandro Conty, Clifford Stein, and Larry Gritz. Sony pictures imageworks arnold. *ACM Transactions on Graphics (TOG)*, 37(3):1–18, 2018. 7

[43] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3587–3596, 2017. 2

[44] Stamatios Lefkimmiatis. Universal denoising networks: a novel cnn architecture for image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3204–3213, 2018. 2

[45] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR*, 2021. 2, 3, 4, 6, 7

[46] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 5

[47] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[48] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *Advances in neural information processing systems*, 31, 2018. 2

[49] Yihao Liu, Anran Liu, Jinjin Gu, Zhipeng Zhang, Wenhao Wu, Yu Qiao, and Chao Dong. Discovering" semantics" in super-resolution networks. *arXiv preprint arXiv:2108.00406*, 2021. 2, 3, 8

[50] Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of super-resolution networks. *arXiv preprint arXiv:2205.07019*, 2022. 3, 8

[51] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *TIP*, 2016. 5

[52] Salma Abdel Magid, Zudi Lin, Donglai Wei, Yulun Zhang, Jinjin Gu, and Hanspeter Pfister. Texture-based error analysis for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2118–2127, 2022. 3

[53] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *2009 IEEE 12th international conference on computer vision*, pages 2272–2279. IEEE, 2009. 2

[54] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29, 2016. 2

[55] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and

measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 4, 5, 7
[56] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3
[57] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1586–1595, 2017. 2
[58] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural information processing systems*, 31, 2018. 2
[59] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3
[60] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3
[61] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 7
[62] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *arXiv preprint arXiv:2207.08494*, 2022. 2
[63] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017. 2
[64] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 5
[65] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010. 3
[66] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 3
[67] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 2
[68] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020. 1, 2
[69] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2, 3
[70] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 2
[71] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 1, 2
[72] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32, 2019. 2
[73] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *European Conference on Computer Vision*, pages 41–58. Springer, 2020. 2
[74] 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 *CVPR*, 2022. 2
[75] 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. 6, 7
[76] 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
[77] Jiale Zhang, Yulun Zhang, Jinjin Gu, Jiahua Dong, Linghe Kong, and Xiaokang Yang. Xformer: Hybrid x-shaped transformer for image denoising. *arXiv preprint arXiv:2303.06440*, 2023. 2
[78] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022. 2, 4
[79] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Radu Timofte, and Luc Van Gool. Practical blind denoising via swin-conv-unet and data synthesis. *arXiv preprint arXiv:2203.13278*, 2022. 2, 5
[80] 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. 1, 2, 6, 7
[81] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 2

[82] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2):023016, 2011. 5

[83] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[84] Yulun Zhang, Kunpeng Li, Kai Li, Gan Sun, Yu Kong, and Yun Fu. Accurate and fast image denoising via attention guided scaling. *IEEE Transactions on Image Processing*, 30:6255–6265, 2021. 2

[85] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2, 6, 7

[86] 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