Efficient Deep Image Prior with Spatial-Channel Attention Transformer

Weiwei Lin *, Zeqing Zhang *, Jin Lin, Ying You

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1060.v1

Keywords: image denoising; efficient deep image prior; spatial-channel attention transformer

 doi:10.20944/preprints202602.1060.v1

Article

Efficient Deep Image Prior with Spatial-Channel Attention Transformer

Weiwei Lin [1,*], Zeqing Zhang [2,*], Jin Lin [3] and Ying You [1]

[1] School of Big Data and Artificial Intelligence, Fujian Polytechnic Normal University, Fuqing City, Fujian Province, China

[2] West Yunnan University of Applied Sciences, Kunming, Yunnan Province, China

[3] Xiamen University, Xiamen, Fujian Province, China

* Correspondence: linww_cn@hotmail.com; 313460472@qq.com

Abstract

The Deep Image Prior(DIP) suggests that it is possible to train a randomly initialized network with a suitable architecture to solve the inverse imaging problem by simply optimizing its parameters to reconstruct a single degraded image. However, the learning effect it seeks is often achieved with the most naive local convolution, which inevitably leads to the inverse imaging problem being limited by the model's generative ability. Furthermore, image info is often not related to surrounding pixels but to overall color and spatial info. Simple local convolution in inverse imaging can't capture precise details. Moreover, DIP is an unsupervised process but requires iterations to learn inverse imaging, consuming computational power and limiting adaption of global attention. To solve these problems, this article explores the possibility of globalizing the DIP task's learning and introducing tri-directional multi-head self-attention to optimize the computation consumption brought by pixel-level attention. Our observations found that global learning can effectively enhance the detail information of edge pixels, making images more vivid and textures clearer. In addition, tri-directional multi-head self-attention can efficiently replace the global perception ability of pixel-level self-attention. Finally, we demonstrate that global learning can effectively improve the imaging effect of inverse imaging problems and enhance the information of texture edge pixels. Moreover, tri-directional multi-head self-attention can effectively alleviate the computation redundancy of pixel-level self-attention, thus achieving efficient and high-quality inverse imaging tasks. The principles of this approach—global feature capture and efficient attention modeling—extend its potential applicability beyond imaging to domains such as software security. For instance, it can enhance tasks like vulnerability analysis by reconstructing obscured code patterns and improve threat modeling through efficient correlation of multi-dimensional attack vectors, balancing detail fidelity with computational practicality.

Keywords: image denoising; efficient deep image prior; spatial-channel attention transformer

1. Introduction

Deep neural networks have been widely used in many computer vision and communication system tasks, and since AlexNe t[17], they have made significant improvements compared to traditional methods. However, image denoising has always been a task where traditional methods (such as BM3D [4]) excelled over early deep learning-based methods [14, 21, 33]. Until the emergence of DnCNN [36], it outperformed traditional methods in dealing with synthetic Gaussian noise at the cost of a large number of noise-free and noisy image pairs[5, 6, 23, 26, 27, 40].

The Deep Image Prior (DIP [28]) model does not require clean and/or noisy image pairs and has demonstrated that a randomly initialized network with an hourglass structure can serve as a prior

 doi:10.20944/preprints202602.1060.v1

for several inverse problems, including denoising, super-resolution[2, 9, 19, 20], and recovering a single degraded image[32, 37, 38]. Although the DIP model exhibits excellent performance in these inverse problems, denoising is a specific task that the DIP model struggles with. Even for synthetic Gaussian noise settings, a single run produces a PSNR that is significantly lower than that of BM 3D.

Additionally, monitoring the PSNR (requiring clean ground truth images) and terminating iteration before fitting the noise is necessary to achieve optimal performance. The deep decoder proposes a strong structure regularization to address this issue, allowing for longer iterations of inverse problems including denoising. However, due to its lower model complexity, its denoising performance is inferior to the DIP model.

However, the existing DIP series of studies still adopt the original ConvNets [3, 11, 13, 15, 16] as backbone for the corresponding unsupervised image recovery task, which inevitably leads to the fact that the recovered results, although capable of smoothing out the noise, will face the common problem shared by localized convolutional neural networks, i.e., unable to ideally deal with the details of the edge texture. Considering that the process of DIP is realized through a certain iterative process, it inevitably consumes a noticeable amount of time to get the desired result. This also leads to the fact that the existing in-depth studies on DIP do not consider enhancing the corresponding image generation capabilities from the most basic aspect of codec networks. This is mainly due to two reasons: 1) the performance of the generative network tends to be positively correlated with its corresponding computational scale, which limits the development of research on improving the generative capability of the model; and 2) the number of iterations of the generative process is huge, so any small increase in the computational efficiency of the generative network will lead to a huge time consumption of the generative process. Therefore, the optimization of this codec backbone must satisfy two conditions at the same time: performance and efficiency.

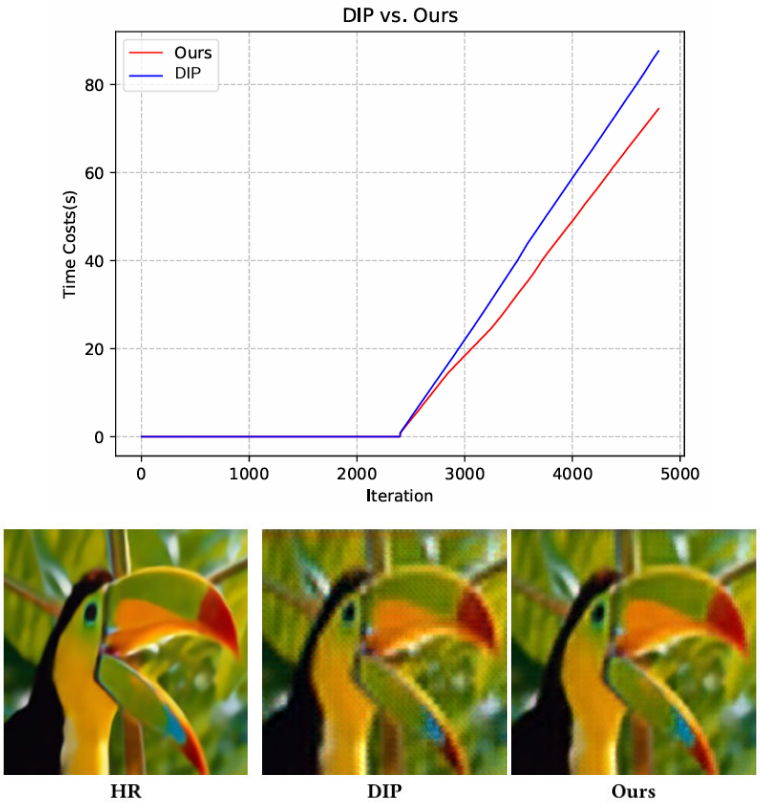

Figure 1. Time costs comparisons and visual comparisons of our methods with DIP on image super-resolution for 4× SR.

 doi:10.20944/preprints202602.1060.v1

In recent years, transformer has developed rapidly in the field of computer vision due to its global retrieval of images. However, the high performance of transformer is often accompanied by a huge computational consumption, and in recent years, high efficiency and high performance transformer-based image restoration methods[29, 32] similar to the Restormer[33] family have emerged.Restormer overcomes the high computational consumption of pixel global attention by calculating channel attention correlation matrices, enabling it to perform image restoration tasks as efficiently as the ConvNets to perform image restoration tasks as efficiently as ConvNets. Inspired by the channel attention mapping, this paper proposes spatial-channel attention for learning global information more efficiently and effectively improving the edge texture details of the generated image without higher than the original computational consumption[5, 6, 26, 34, 35, 39]. Among them, spatial-channel self-attention additionally adds the learning of self-attention maps in the horizontal and vertical spaces in the image to the learning of the original channel-level attention maps. For pixel-level global attention, the learned relevant information is not always appropriate. Moreover, pixel-level global attention learning requires very large hardware space requirements as well as computational consumption and time consumption. Therefore, spatial-channel self-attention can effectively avoid computational redundancy and improve imaging performance. We summarize the main contributions of this work as follows:

- We propose an unsupervised DIP method based on TMTA. Our method is optimized on the backbone of the original DIP, which means that all methods for subsequent DIP sequences can seamlessly incorporate our optimization. Thus, it can directly improve the performance of the methods related to the DIP series.
- We propose for the first time a DIP approach based on self-attention. Our approach is competitive with ConvNets in terms of efficiency, while a significant improvement in recovery performance makes it applicable to the iterative generation process of DIP.
- As shown in Fig. 1, extensive experiments demonstrate that our approach has demonstrable improvements in performance and time consumption. Moreover, the method can be used as an optimized version of the subsequent DIP algorithm.

2. Proposed Method

This section will introduce a efficient DIP (TM-DIP) based on Triple Multi-Head Transposed Attention (TMTA). Our main goal is to introduce global self-attention into the DIP task. Unlike previous DIP methods, TM-DIP overcomes the global smoothness of the convolution process from the computational efficiency point of view, and enhances the information of the edge texture details through global self-attention learning. As a result, our method achieves an organic unity in efficiency and performance. In the following subsections, we will first give an overview of the proposed TM-DIP. Then, we analyze the computational efficiency of the proposed global attention of the method. finally, we will elaborate on the core components of the proposed method.

2.1. Deep Image Prior (DIP)

Let a noisy image $y \in \mathbb{R}^N$ be modeled as:

$$y = x + n, \tag{1}$$

where $x \in \mathbb{R}^N$ be a noiseless image that one would like to recover and $n \in \mathbb{R}^N$ be an *i.i.d.* Gaussian noise such that $n \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is an identity matrix. Denoising can be formulated as a problem of predicting the unknown x from known noisy observation y. Ulyanov *et al.* [27, 28] argued that a network architecture naturally encourages to restore the original image from a degraded image y and name it as deep image prior (DIP). Specifically, DIP optimizes a convolutional neural network h with parameter by a simple least square loss $\mathcal{L}$ as:

$$\hat{\theta} = \arg\min \mathcal{L}(\mathbf{h}(\hat{n};\theta), y), \tag{2}$$

 doi:10.20944/preprints202602.1060.v1

θ

where $\hat{n}$ is a random variable that is independent of y. If **h(•)** has enough capacity (*i.e.*, sufficiently large number of parameters or architecture size) to fit to the noisy image y, the output of model h($\hat{n};\theta$) should be equal to y, which is not desirable. DIP uses the early stopping to obtained the results with best PSNR with clean images.

2.2. Backbone of Deep Image Prior (DIP)

The prior defined by implicit Eq. 2 is implicit and does not define an appropriate probability distribution in image space. However, it is possible to extract "samples" (in a loose sense) from this prior by means of random values of the parameter θ and the generated image $f(\theta)$. In other words, we can visualize the starting point of the optimization process Eq. 2, before fitting the parameters to the noisy image.DIP analyzes such "samples" from the depth prior captured by different hourglass-type architectures. The generative network chosen for the classical DIP is still used in subsequent studies as shown in Fig. 2. Therefore, this architecture is naturally the most popular choice for generating ConvNets[7, 8, 30, 31]. Instead, we optimize the generative network of the DIP from its roots so that it can be applied to all DIP-based research efforts.

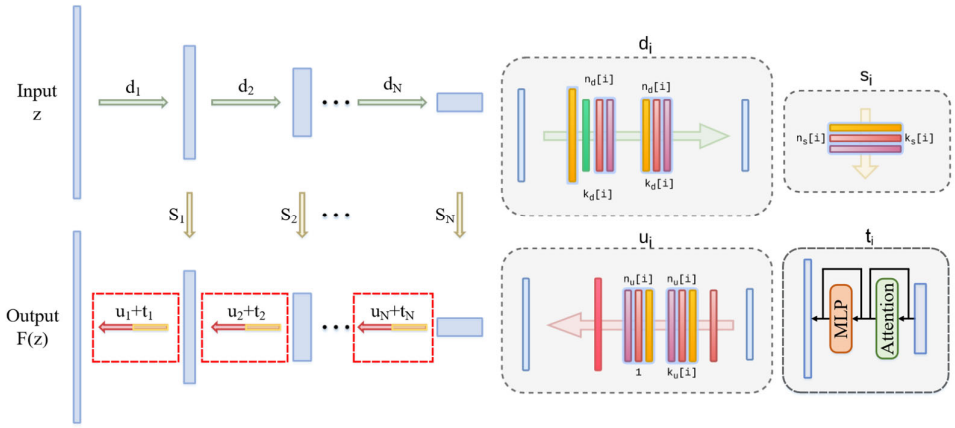

Figure 2. Illustration of our proposed TM-DIP. We use "hourglass" (also known as "decoder-encoder") architecture same as the classical DIP. We sometimes add skip connections (yellow arrows). u_i, d_i, s_i, t_i correspond to the number of filters at depth i for the upsampling, downsampling, skip-connections and non-local selfattention respectively. $n_u[i]$, $n_d[i]$, $n_s[i]$ correspond to the number of filters at depth i for the upsampling, downsampling and skip-connections respectively. The values $k_u[i]$, $k_d[i]$, $k_s[i]$ correspond to the respective kernel sizes.

2.3. Overview of TM-DIP

 doi:10.20944/preprints202602.1060.v1

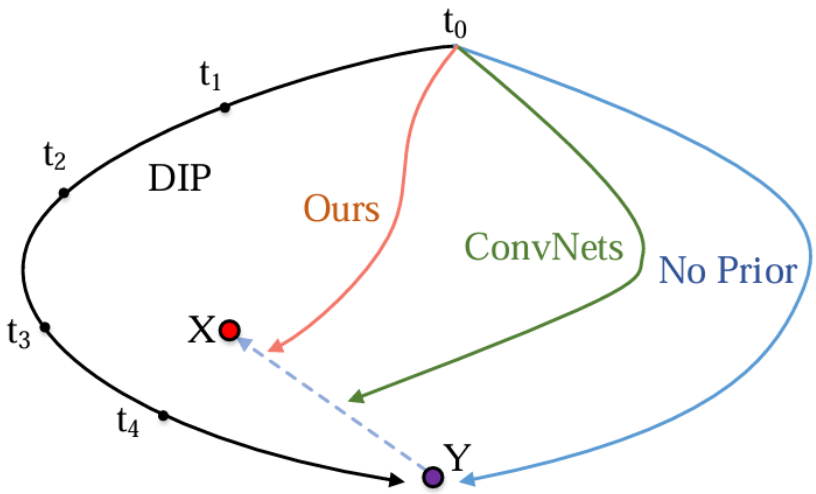

Figure 3. Illustration of a solution trajectory of ours and DIP. We consider the problem of reconstructing an image X from a degraded measurement Y. Ours changes DIP's solution trajectory from black to orange and is close to noiseless solution (X).The black line denotes the ideal state optimized by iteration to the degenerate image Y. The purple line denotes the direct mapping without adding any prior knowledge. The green line denotes the optimization path with ConvNets as the baseline, representing local optimization as well as global homogenization. The brown line denotes our transformer-based optimization path, representing global optimization as well as local edge texture enhancement.

As shown in Fig. 2, there are four main components: the downsampling module, the upsampling module, the hopping connection module, and our proposed spatial-channel self-attention module. Among them, the down-sampling module, the up-sampling module, and the hopping connection still follow the structure of DIP, while the added spatial-channel self-attention module can be embedded into the generative network of DIP as a plug-and-play module. As shown in Fig. 2, the red dashed box shows the process of combining the up-sampling and spatial-channel self-attention modules. Although ConvNets are able to accomplish the task of high-fidelity imaging of the original image, they are overly noise-smoothed due to local convolutions and are homogeneous with respect to the global information, mainly because most of the regions in the image are smooth and unvarying. As shown in Fig. 1, the defects of DIP at the edge texture sites are obvious, while the optimization of TM-DIP at the edge sites is significant. In other words, it is understood that Y is the degraded texture image and X is the corresponding clear image as shown in Fig. 3. ConvNets-based DIP is able to smooth out the degraded part of the Y image, but it also smooths out the important edge texture information and content in X. The DIP can be used to optimize the edge texture information and content in X, while the TM-DIP can be used to optimize the edge texture information. On the other hand, TM-DIP based on spatial-temporal global attention is committed to add global attention on top of convolutional smoothing of noise, which greatly recovers the important edge texture information and content in X. In addition, the effect of convolutional smoothing makes TM-DIP's recovery ability somewhat limited, i.e., it does not have the ability to recover degraded texture information (degraded texture information is independently uncorrelated with image information). It is this one characteristic that TM-DIP is just in the interval of recovering edge texture information and degradation information, making it only able to recover detail information but not able to pay degradation information, which is in line with the optimal state of the image recovery task.

2.4. Triple Multi-Head Transposed Attention

In addition to the above approaches, TMTA considers a very significant issue: the limitation of Transformer in image restoration lies in the huge computational complexity caused by the demand to complete high-resolution correlation calculation between various pixels. As shown in Fig. 2, the

 doi:10.20944/preprints202602.1060.v1

pixel magnification of intermediate features has been scaled by ×16 , and the computation can be reduced by ×256 if the traditional self-attention mechanism of full pixels is adopted. Even so, when it comes to higher resolution images, there is still the problem of "high computational complexity". To this end, considering the information redundancy of full-pixel self-attention, TM-DIP proposes Triple Multi-Head Transposed Attention (TMTA). It decomposes the attention of characteristic pixels into three directions of self-attention for cooperative computation: horizontal self-attention, vertical self-attention, and channel self-attention.

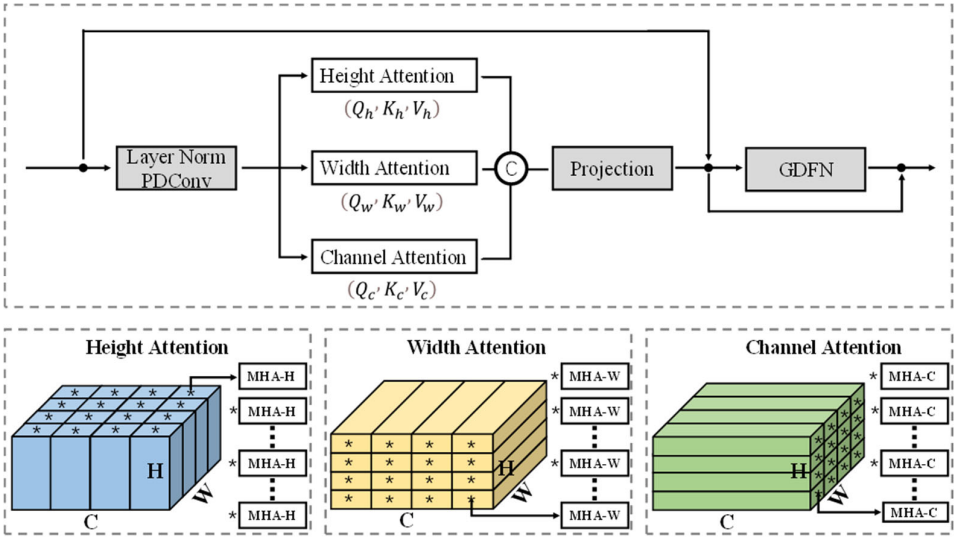

Figure 4. Illustration of Triple Multi-Head Transposed Attention module(TMTA). The attention of characteristic pixels is decomposed into three directions of self-attention for cooperative computation: horizontal self-attention, vertical self-attention, and channel self-attention.

As shown in Fig. 4, the input features first pass through the "Layer Norm+PDConv" layer to generate the locally enriched *query*(Q), *key*(K) and *value*(V). The Layer Norm (LN) denotes the regular layer normalization, and the PDConv denotes the combination of Pointwise Convolution (PWConv) and Depthwise Convolution (DWConv). Then, the *query*(Q) and *key*(K) are reshaped in three-dimensional directions, resulting in the horizontal $query_H$ (Q_H) and key_H (K_H), the vertical $query_W$ (Q_W) and key_W (K_W), and the $query_C$(Q_C) and key_C(K_C), respectively. Then, matrix multiplication is performed on them respectively to generate three transposed attention matrices(Fig. 5) with sizes of $\mathbb{R}^{H\times H}$, $\mathbb{R}^{W\times W}$ and $\mathbb{R}^{C\times C}$, instead of the regular attention matrix $\mathbb{R}^{HW\times HW}$ of characteristic pixels [10, 29]. It is worth noting that all three processes are transformed from *query*(Q), *key*(K) and are synergistically related to each other. In general, the process definition of TMTA is as follows:

$$
\begin{aligned}
X' &= W_p\,\mathrm{Atention}(Qs, Ks, Ys) + X,\\
\mathrm{Atention}(Qs, Ks, Vs) &= \mathrm{Concat}(A_H, A_W, A_C),\\
A_H &= V_H \times \mathrm{Softmax}(K_H \times Q_H/\alpha_H),\\
A_W &= V_W \times \mathrm{Softmax}(K_W \times Q_W/\alpha_W)\\
A_C &= V_C \times \mathrm{Softmax}(K_C \times Q_C/\alpha_C)
\end{aligned}
\tag{3}
$$

where X and X′ denote the input and output features; $Q_i \in (\mathbb{R}^{WC\times H}, \mathbb{R}^{HC\times W}, \mathbb{R}^{HW\times C})$, $K_i \in (\mathbb{R}^{H\times WC}, \mathbb{R}^{W\times HC}, \mathbb{R}^{C\times HW})$, $V_i \in (\mathbb{R}^{WC\times H}, \mathbb{R}^{HC\times W}, \mathbb{R}^{HW\times C})$ denotes the horizontal, vertical, and channel reshaping by the generated*query*(Q), *key*(K) and *value*(V), respectively; α_i denotes a learnable

 doi:10.20944/preprints202602.1060.v1

scaling parameter to control the size of the dot product of Q_i and K_i before applying the activation function. In the above expression, $i \in$ [H, W, C].

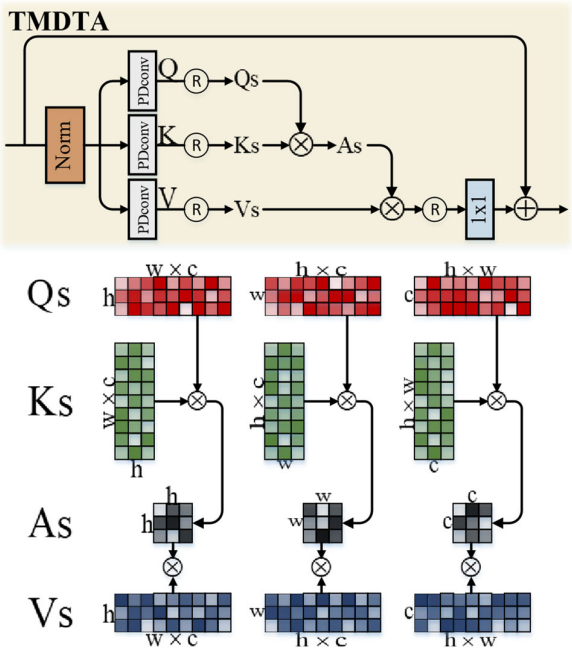

Figure 5. Illustration of Attention module of Triple Multi-Head Transposed Attention. TMTA's attention consists of a LayerNorm module, three depth separable convolutions (PDConv), six matrix multiplications, and three point-wise convolutions.

2.5. Efficiency of TMTA

In this subsection, we analyze why our proposed TMTA possesses high efficiency. First, assume that the current shape of the input features is (B, C, H, W), the size of the convolution kernel is (K, K, C, C), and the head of the multi-head attention is set to 1 for ease of computation (the value of h does not affect the comparison results). The traditional ViT calculates the correlation between each pixel of an image, so its computation is positively correlated with the pixels of the image. With the continuous iteration and update of the technology, the requirements for image processing are now gradually increased from low resolution to high resolution. Therefore, its self-attention is calculated as the matrix product between the QUERY of shape (B, H × W, C) and the KEY of shape (B, C, H × W) to get the attention map of shape (B, H × W, H × W). The computation of this part is:

$$OP(\text{ViT}) = B \times H \times W \times H \times W \times C. \quad (4)$$

And the self-attention of TM-DIP is computed as the matrix product between QUERY of shape (B, C, H × W) and KEY of shape (B, H × W, C) to get the attention mapping of shape (B, C, C). Its corresponding computation is:

$$OP(\text{TMTA}) = B \times C \times H \times W \times C. \quad (5)$$

It can be seen that the difference between the two equations lies in H × W and C. For an image of common resolution such as 256 × 256, the value of channel C hardly exceeds 512, which can be seen to be much smaller than H × W. Therefore, in terms of computation and memory requirements, TM-DIP has a huge improvement in computational efficiency compared to all-pixel self-attention ViT, while in terms of performance it can sufficiently overcome the computation of redundancy, which is mainly due to the extremely high percentage of homogeneous pixels (background) for images.

3. Experiments

 doi:10.20944/preprints202602.1060.v1

3.1. Experimental Setup

Implementation Details. To ensure the fairness of the comparison between methods, our method and classical DIP methods adopt the same classic datasets, include denoising, super-resolution, flash reconstruction and inpainting [1]. Consistent with the most primitive DIP, we used an encoder-decoder ("hourglass") architecture (possibly with skip connections) for f in all experiments unless otherwise noted in Fig. 2, varying the hyperparameters by a small amount.

Evaluation metrics. We use peak signal-to-noise ratio (PSNR) and cost time(s). The PSNR is widely used in denoising literature [16, 29, 51, 52, 53] but is recently argued that it is not an ideal metric as it values the oversmoothed results [21, 54]. We use the publicly available pre-trained weights based on AlexNet by the authors [54]. We additionally report the performance of the peak PSNR during optimization of our method as a reference (denoted as «Ours*»).

3.2. Comparison with DIP on Denoising and Generic Reconstruction

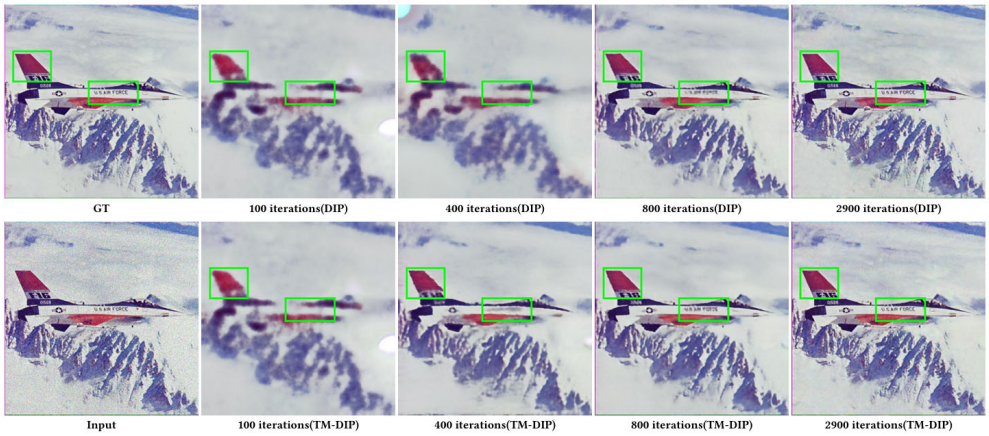

Figure 6. Blind image denoising. The deep image prior is successful at recovering both man-made and natural patterns. TM-DIP is able to recover detailed information on the eve of the number of iterations, whereas DIP can only approximate the result from a slow global averaging, which leads to deficiencies in important detail areas.

Our approach is consistent with the original DIP in that it does not model the image degradation process that it requires for recovery. This allows it to be applied in a "plug-and-play" manner to DIP-based image restoration tasks, where the degradation process is complex and/or unknown, and real data for supervised training is difficult to obtain. We validate that TM-DIP is effective and outperforms DIP in detailed areas by using the qualitative example in Fig. 6. As can be seen in the figure, TM-DIP is able to quickly focus on the learning of detailed regions at each stage of the iteration. For the iterative generation stages of DIP and TM-DIP, it can be observed that the convolutional neural network focuses on learning from global homogeneity, while the introduction of self-attention effectively enhances the detail information as well as the clarity. In other words, self-attention focuses on the approximation of edge details, while the convolutional neural network test focuses on the smoothing of overall accuracy.

 doi:10.20944/preprints202602.1060.v1

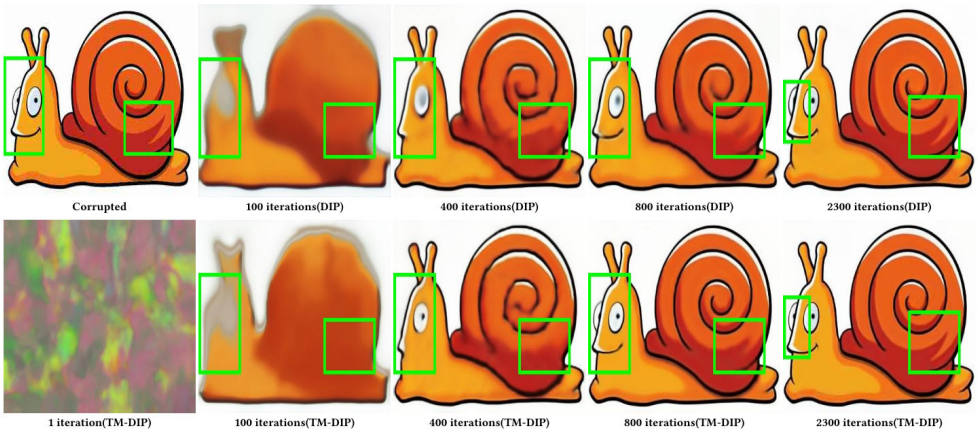

Figure 7. Blind restoration of a JPEG-compressed image. (electronic zoom-in recommended). Our approach can restore an image with a complex degradation (JPEG compression in this case).

Fig. 7 also similarly demonstrates the learning efficiency of TM-DIP for detail information. TMDIP meaty is able to pay stronger attention to detail regions, while convolution-based DIP is based on global averaging for optimal learning, and the final result on DIP is blurred by the meaty in the edge texture region.

3.3. Comparison with DIP on Super-resolution

Table 1. Detailed super-resolution PSNR comparison on the 4× Set14.

	Baboon	Barbara	Bridge	Coastguard	Comic	Face	Flowers	Foreman	Lena	Man	Monarch	Pepper	Ppt3	Zebra	Avg
No prior	22.24	24.89	23.94	24.62	21.06	29.99	23.75	29.01	28.23	24.84	25.76	28.71	20.26	21.69	24.93
Bicubic	22.44	24.15	24.47	25.53	21.59	31.34	25.33	29.45	29.84	25.7	27.45	30.63	21.78	24.01	26.05
TV prior [22]	22.34	24.78	24.46	25.78	21.95	31.34	25.91	30.63	29.76	25.94	28.46	31.32	22.75	24.52	26.42
Glasner et al.[12]	22.44	25.38	24.73	25.38	21.98	31.09	25.54	30.4	30.48	26.33	28.22	32.02	22.16	24.34	26.46
DIP	22.29	25.53	24.38	25.81	22.18	31.02	26.14	31.66	30.83	26.09	29.98	32.08	24.38	25.71	27.00
Ours	22.31	25.63	24.45	25.94	22.29	31.17	26.28	31.73	30.99	26.14	30.12	32.21	24.43	25.85	27.11
SRResNet-MSE [19]	23.00	26.08	25.52	26.31	23.44	32.71	28.13	33.8	32.42	27.43	32.82	34.28	26.56	26.95	28.53
LapSRN [18]	22.83	25.69	25.36	26.21	22.9	32.62	27.54	33.59	31.98	27.27	31.62	33.88	25.36	26.98	28.13

Table 2. Detailed super-resolution PSNR comparison on the 8× Set14.

	Baboon	Barbara	Bridge	Coastguard	Comic	Face	Flowers	Foreman	Lena	Man	Monarch	Pepper	Ppt3	Zebra	Avg

 doi:10.20944/preprints202602.1060.v1

No prior	21.09	23.04	21.78	23.63	18.65	27.84	21.05	25.62	25.42	22.54	22.91	25.34	18.15	18.85	22.56
Bicubic	21.28	23.44	22.24	23.65	19.25	28.79	22.06	25.37	26.27	23.06	23.18	26.55	18.62	19.59	23.09
TV prior[22]	21.3	23.72	22.3	23.82	19.5	28.84	22.5	26.07	26.74	23.53	23.71	27.56	19.34	19.89	23.48
SelfExSR[25]	21.37	23.9	22.28	24.17	19.79	29.48	22.93	27.01	27.72	23.83	24.02	28.63	20.09	20.25	23.96
DIP	21.38	23.94	22.2	24.21	19.86	29.52	22.86	27.87	27.93	23.57	24.86	29.18	20.12	20.62	24.15
ours	21.49	24.07	22.31	24.42	19.97	29.71	22.95	27.94	28.06	23.75	24.98	29.31	20.15	20.71	24.37
LapSRN [18]	21.51	24.21	22.77	24.10	20.06	29.85	23.31	28.13	28.22	24.20	24.97	29.22	20.13	20.28	24.35

Table 3. Detailed super-resolution PSNR comparison on the 4× Set5.

	Baby	**Bird**	**Butterfly**	**Head**	**Woman**	**Avg**
No prior	30.16	27.67	19.82	29.98	25.18	26.56
Bicubic	31.78	30.2	22.13	31.34	26.75	28.44
TV prior[22]	31.21	30.43	24.38	31.34	26.93	28.85
SelfExSR [25]	32.24	31.1	22.36	31.69	26.85	28.84
DIP	31.49	31.8	26.23	31.04	28.93	29.89
ours	32.25	31.95	26.45	31.17	29.21	30.21
LapSRN [18]	33.55	33.76	27.28	32.62	30.72	31.58
SRResNet-MSE [19]	33.66	35.1	28.41	32.73	30.6	32.1

Table 4. Detailed super-resolution PSNR comparison on the 8× Set5.

	Baby	**Bird**	**Butterfly**	**Head**	**Woman**	**Avg**
No prior	26.28	24.03	17.64	27.94	21.37	23.45
Bicubic	27.28	25.28	17.74	28.82	22.74	24.37
TV prior[22]	27.93	25.82	18.4	28.87	23.36	24.87
SelfExSR [25]	28.45	26.48	18.8	29.36	24.05	25.42
DIP	28.28	27.09	20.02	29.55	24.5	25.88
ours	28.41	27.22	20.13	29.71	24.67	26.05
LapSRN [18]	28.88	27.1	19.97	29.76	24.79	26.1

We similarly use the center crop of the generated image to compute the PSNR (Tables 1, 2, 3 and 4). Our method, while outperforming DIP in terms of accuracy, is still lower than the learning-based method. However, learning-based methods require a large amount of training time, which is a common problem with supervised learning. As can be seen from the tables, the improvement of TM-DIP is obvious.

3.4. Comparison with DIP on Inpainting

 doi:10.20944/preprints202602.1060.v1

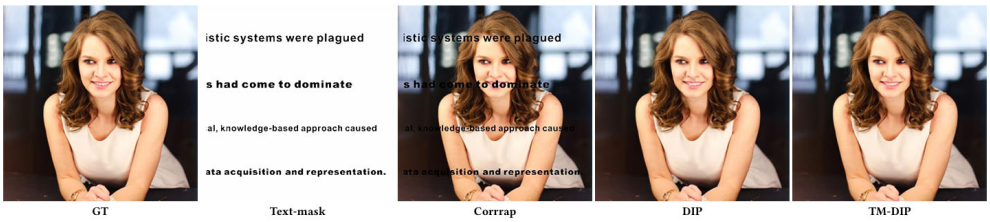

Figure 8. Comparison on text the inpainting task.

Figure 9. Inpainting using different depths.

We similarly compare the results in terms of inpaiting. The results were similar in terms of visual sensations for text inpainting(Fig. 8). In Fig 9, we compare the depth prior corresponding to different levels of architecture. We compare the visual differences between TM-DIP and DIP between different levels of architecture. It can be seen that the results recovered by TM-DIP still focus on non-smooth regions, while the results recovered by DIP focus more on smooth regions.

3.5. Comparison with DIP on Flash-no Flash Reconstruction

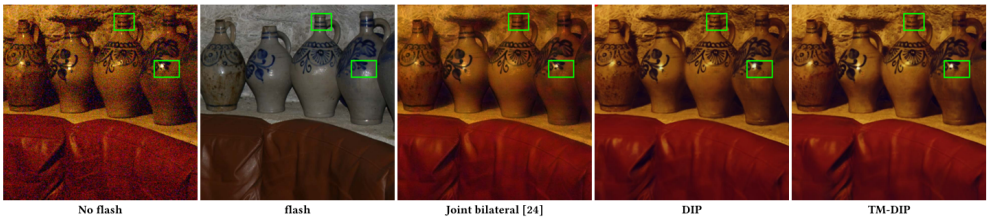

Figure 10. Reconstruction based on flash and no-flash image pair. The deep image prior allows to obtain low-noise reconstruction with the lighting very close to the no-flash image. It is more successful at avoiding "leaks" of the lighting patterns from the flash pair than joint bilateral filtering [24] (c.f. blue inset).

This subsection shows the results of the comparison in terms of Flash-no Flash Reconstruction in Fig. 10. It can be observed that the results of TM-DIP over DIP are closer to the original image in terms of color tone and detail. Moreover, the imaging results of DIP are diffuse and cannot have a proper contrast, or the contrast is more uniform. In contrast, MT-DIP is able to focus on the enhancement of specific areas as well as adapt the overall contrast. In addition, the global averaging learning of DIP makes the overall color tone of the image less a priori than TM-DIP.

 doi:10.20944/preprints202602.1060.v1

3.6. Comparison with DIP on Time cost

As shown in Fig. 2, TM-DIP implements global self-attention learning by replacing the convolutional structure in DIP. As previously analyzed, the convolutional structure of DIP contains multiple base operations instead of a single convolutional process, while the structure of our proposed TMTA is simple, and the computational complexity approximates that of a single convolution. In particular, it should be noted that TMTA and the conventional transformer have significant advantages in the processing of high-resolution images. This is due to the fact that TMTA is proportional to the image resolution, whereas the conventional transformer is proportional to the square of the image resolution, and this relationship creates an unbridgeable gap with increasing resolution. As shown in Fig. 1, the computational efficiency of TM-DIP is even better than that of DIP, which fully validates the computational efficiency of our proposed method as well as its foresight.

4. Conclusions

In this paper, we propose an efficient image denoising baseline called TM-DIP.TM-DIP introduces the Triple Multi-Head Transposed Attention mechanism, TMTA, in DIP. The TMTA mechanism decomposes the traditional all-pixel self-attention mapping computation into horizontal self-attention, vertical self-attention, and channel self-attention. Among them, the horizontal and vertical self-attention fit the spatial information of features, so TMTA can fully learn the feature information in the spatial-channel dimension. In addition, TMTA realizes the effect of fast computation by this multi-directional disassembly, and has improved time efficiency over DIP, while the visual effect is significantly improved over DIP in the learning of detail edges. In addition, the advantage of TMTA is that it can be applied to any other architectures as a standalone module with a certain guarantee of computational efficiency. And for all subsequent DIP methods, the addition of TMTA can provide a stronger enhancement to their findings. The principles of global learning and efficient structured attention presented in this work can also facilitate the reconstruction and analysis of complex structured data from corrupted inputs, with direct applicability to domains such as software security for vulnerability analysis.

Funding: This work is supported by the Natural Science Foundation of Fujian Province, China (No. 2023J011119); the Natural Science Foundation of Fujian Province, China (No. 2023J011118); and 2025 Yunnan Provincial Department of Education Scientific Research Fund Project, China (2025J1100).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. 2018. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1692–1700.
2. Saeed Anwar, Salman Khan, and Nick Barnes. 2020. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.
3. Andrea Asperti and Valerio Tonelli. 2023. Comparing the latent space of generative models. *Neural Computing and Applications* 35, 4 (2023), 3155–3172.
4. Harold C Burger, Christian J Schuler, and Stefan Harmeling. 2012. Image denoising: Can plain neural networks compete with BM3D. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2392–2399.
5. Yunjin Chen and Thomas Pock. 2016. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1256–1272.
6. Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. 2021. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4896–4906.

 doi:10.20944/preprints202602.1060.v1

7. Xuan Ding, Hongchao Fan, and Jianya Gong. 2021. Towards generating network of bikeways from Mapillary data. *Computers, Environment and Urban Systems* 88 (2021), 101632.
8. Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13733–13742.
9. Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2016. Image super-resolution using deep convolutional networks. 38, 2 (2016), 295–307.
10. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
11. Alexey Dosovitskiy, Thomas Brox, et al. 2015. Inverting convolutional networks with convolutional networks. *arXiv preprint arXiv:1506.02753* 4, 2 (2015), 3.
12. Daniel Glasner, Shai Bagon, and Michal Irani. 2009. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*. IEEE, 349–356.
13. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
14. Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2019. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1712–1722.
15. Xinhui Kang, Shin'ya Nagasawa, Yixiang Wu, and Xingfu Xiong. 2023. Emotional design of bamboo chair based on deep convolution neural network and deep convolution generative adversarial network. *Journal of Intelligent & Fuzzy Systems* Preprint (2023), 1–13.
16. Farhan Khawar, Leonard Poon, and Nevin L Zhang. 2020. Learning the structure of auto-encoding recommenders. In *Proceedings of The Web Conference 2020*. 519–529.
17. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).
18. Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 624–632.
19. Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2019. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*. 4681–4690.
20. Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPRW*. 136–144.
21. Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. 2021. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13365–13374.
22. Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5188–5196.
23. Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems* 29 (2016).
24. Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. 2004.Digital photography with flash and no-flash image pairs. *ACM transactions on graphics (TOG)* 23, 3 (2004), 664–672.
25. Assaf Shocher, Nadav Cohen, and Michal Irani. 2018. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3118–3126.
26. Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. 2017. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*. 4539–4547.

 doi:10.20944/preprints202602.1060.v1

27. Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9446–9454.
28. Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9446–9454.
29. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
30. Shiping Wen, Weiwei Liu, Yin Yang, Tingwen Huang, and Zhigang Zeng. 2018. Generating realistic videos from keyframes with concatenated GANs. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 8 (2018), 2337–2348.
31. Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. 2019. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10734–10742.
32. Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. 2022. Efficient Non-Local Contrastive Attention for Image Super-Resolution. *AAAI* (2022).
33. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5728–5739.
34. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2020. 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*. Springer, 492–511.
35. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14821–14831.
36. Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing* 26, 7 (2017), 3142–3155.
37. Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*. 286–301.
38. Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image superresolution. In *CVPR*. 2472–2481.
39. Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2020. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 7 (2020), 2480–2495.
40. Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. 2020. When awgnbased denoiser meets real noises. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13074–13081.