Generative Adaptive Convolutions for Real-World Noisy Image Denoising

Ruijun Ma[1, 2], **Shuyi Li**[1], **Bob Zhang**[1]*, **Zhengming Li**[2]

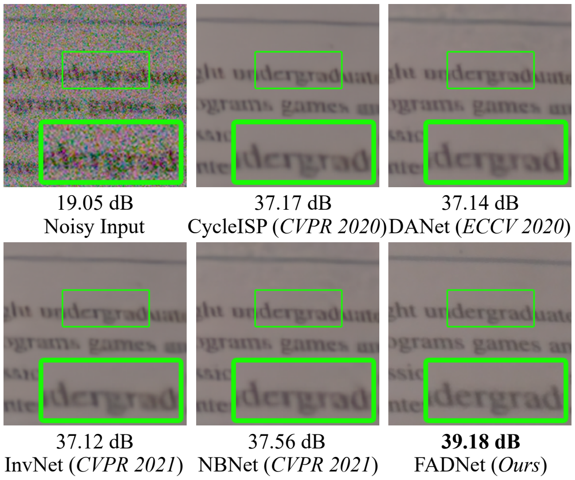
[1] PAMI Research Group, Department of Computer and Information Science, University of Macau
[2] Guangdong Industrial Training Center, Guangdong Polytechnic Normal University
{yb97442, yb97443, bobzhang}@um.edu.mo, gslzm@gpnu.edu.cn

Abstract

Recently, deep learning techniques are soaring and have shown dramatic improvements in real-world noisy image denoising. However, the statistics of real noise generally vary with different camera sensors and in-camera signal processing pipelines. This will induce problems of most deep denoisers for the overfitting or degrading performance due to the noise discrepancy between the training and test sets. To remedy this issue, we propose a novel **f**lexible and **a**daptive **d**enoising **net**work, coined as FADNet. Our FADNet is equipped with a plane dynamic filter module, which generates weight filters with flexibility that can adapt to the specific input and thereby impedes the FADNet from overfitting to the training data. Specifically, we exploit the advantage of the spatial and channel attention, and utilize this to devise a decoupling filter generation scheme. The generated filters are conditioned on the input and collaboratively applied to the decoded features for representation capability enhancement. We additionally introduce the Fourier transform and its inverse to guide the predicted weight filters to adapt to the noisy input with respect to the image contents. Experimental results demonstrate the superior denoising performances of the proposed FADNet versus the state-of-the-art. In contrast to the existing deep denoisers, our FADNet is not only flexible and efficient, but also exhibits a compelling generalization capability, enjoying tremendous potential for practical usage.

Introduction

Image denoising is a fundamental computer vision task, aiming to recover the latent clean image from its counterpart noisy observation. In the past years, a vast amount of denoising methods have been proposed and obtained a continuous performance growth. Most of these methods are dedicated to additive white Gaussian noise (AWGN) removal and have achieved near-optimal performances (Zhang et al. 2017; Zhang, Zuo, and Zhang 2018; Jia et al. 2019; Xu et al. 2021; Zhang et al. 2021). Whereas in real CCD or CMOS camera systems, the image noises are heavily transformed by the in-camera signal processing (ISP) pipeline and contain multiple different sources (Tsin, Ramesh, and Kanade 2001; Kim et al. 2012). The statistics of real noise, oftentimes, are signal-dependent, spatially variant and do not necessarily remain uniform like AWGN (Seybold et al. 2014; Karaimer and Brown 2016). The assumption that the AWGN and real noise follow the same distribution, however, may limit the applicability greatly on real-world denoising tasks.


19.05 dB
Noisy Input

37.17 dB
CycleISP (*CVPR 2020*)

37.14 dB
DANet (*ECCV 2020*)

37.12 dB
InvNet (*CVPR 2021*)

37.56 dB
NBNet (*CVPR 2021*)

39.18 dB
FADNet (*Ours*)

Figure 1: Denoising example from the Smartphone Image Denoising Dataset (SIDD). Compared to the recent state-of-the-art methods, our algorithm can better preserve the structural contents and image textures, while eliminating the complex real noise. Please view in color with zoom.

Recent research on real-world noisy image denoising has progressed dramatically with the rapid advance of convolutional neural networks (CNNs). The boost in performance of most deep denoisers is overly dependent on whether the domains of the training and test sets are well matched. Actually, the statistics of real noise are device-dependent. Factors such as camera sensors, camera settings, ISP pipelines and scenes can cause heavy noise discrepancy among real camera images. Thus, simply training the deep denoisers on the open benchmark datasets (i.e., SIDD (Abdelhamed, Lin, and Brown 2018) and Nam (Nam et al. 2016)) will deprive their generalization ability to adapt to a wider range of scenarios.

As a remedy, (Guo et al. 2019; Kim et al. 2020) provided an alternative solution by modeling the noise distribution. These methods simulate the response functions of the ISP pipelines to establish a more generalized noise model. One

*Corresponding Author.

clear advantage is that they can generate noisy images with distinct characteristics. Thus, the real denoisers can be readily trained to handle a large pool of noisy images with unknown noise types. In this sense, the overfitting issue of the aforementioned methods can be well solved. However, these approaches usually overlook the influence of the noise properties from the camera sensors, and simplify the reverse engineer of the ISP pipelines, casting the flexibility of the noise modeling-based methods with respect to practical applications into doubt.

In this paper, we tackle the above issues and limitations by presenting a flexible and adaptive denoising network (FADNet) with our novel plane dynamic filter module (PDFM). PDFM is specifically designed for endowing the FADNet with the ability to generalize well on the unknown and sophisticated real noise. To achieve this, PDFM employs a dynamic filter prediction scheme, in which the weight filters are generated conditioned on the input and applied to the extracted deep features. Moreover, inspired by the attention mechanisms, we propose to decouple the filter generation procedure into spatial and channel pathways. As such, the channel and spatial filters can be interwoven within a unified module and the flexibility of the adaptive feature learning can be further enhanced. To refine the predicted filters with rich contextual information, we seek to introduce the Fourier transform and make it guide the PDFM to exploit better the semantic and color features. As shown in Fig. 1, compared with the recently leading denoising methods, our FADNet achieves very pleasing results by preserving clear structures and details even under severe noise.

Summarily, our contributions include:

(1) We propose a novel and flexible denoiser, which is well-generalized and can work universally well for test images with noise discrepancy.

(2) We incorporate the Fourier transform into the dynamic filter generation step for adaptive representation capability enhancement. To the best of our knowledge, this scheme is the pioneering attempt towards merging the power of dynamic filter and Fourier transform for real noise removal.

(3) Comprehensive experiments show that our FADNet enjoys appealing efficiency, whilst achieving a higher performance than the state-of-the-art image denoising methods.

Related Work

Real Image Denoising

In recent years, deep CNNs have made a splash in the arena of real noise removal. Driven by the great capabilities of the deep networks and the ease of access to external training data, the CNN-based denoisers have shown dramatic improvements and exhibited a large performance leap. Zhang et al. (Zhang et al. 2017) proposed to apply the residual learning and batch normalization to facilitate the CNN training for blind image denoising. In (Zhang, Zuo, and Zhang 2018), the authors leveraged the benefits of a tunable noise level map to recover the corrupted noisy images. Yue et al. (Yue et al. 2019, 2020) introduced a generative framework, where the tasks of noise removal and generation can be simultaneously attained within a unique Bayesian model. Ma et al. (Ma et al. 2020) incorporated a set of Kalman filters into a pyramid neural network to remove the noise in a coarse-to-fine manner. Recently, visual attention mechanism has exhibited blossoming developments in the field of image denoising (Saeed and Nick 2019; Ma et al. 2021a,b; Syed et al. 2020). For instance, (Syed et al. 2020) applied the dual-attention units to encode multi-scale context for spatially-precise representations, enabling the denoised results to maintain high-resolution details. Just recently, (Liu et al. 2021) proposed a lightweight model by exploring the invertible networks. This model used two different latent variables to eliminate and generate noise jointly. The algorithm (Cheng et al. 2021) benefitted from the image-adaptive projection technology and had shown promising performance on synthetic and real noise removal.

Albeit the rapid advance of the above CNN-based methods, they could easily overfit to specific training data and lack in generalizability on test images with noise discrepancy. An alternative was to simulate the generation procedure of the ISP pipeline (Guo et al. 2019; Kim et al. 2020). Nonetheless, simply reversing the response functions of the ISP pipeline was not sufficient to model the full characteristics of the real noise. In stark contrast to the above, our FADNet armed with the dynamic filters can cope with the real noise soundly and feasibly, while exhibiting superior generalization ability on test images with unknown noises.

Dynamic Filters

Different from the traditional CNN, where the filter weights stay fixed once trained, the dynamic filters are generated by separate network branches and can be changed according to the input on-the-fly. Due to its adaptive nature, it can increase the flexibility of a network and has been applied to various tasks, such as super-resolution (Hu et al. 2019), point cloud segmentation (Xu et al. 2020), image deblurring (Lee et al. 2021), and style transfer (Chandran et al. 2021). However, generating such depthwise-separable and spatially-varying filters usually entailed memory intensive network architectures, which was computation-heavy and time-consuming.

Our work circumvented the above dilemmas through decoupling the generated filters into spatial and channel ones. Closest to this scheme was the method proposed in (Zhou et al. 2021). However, (Zhou et al. 2021) generated filters at each channel dimension. Conversely, we emphasize that our proposed model is more lightweight in that we predict the weight kernels along the channel dimension. Moreover, we share the generated filters over channels for feature encoding, such that the spatial and channel information can be interwoven for more informative representations.

Proposed Method

Network Architecture

Fig. 2 illustrates an overview of the proposed denoising network, namely FADNet. Our FADNet takes a noisy image I_N as input and generates a denoised output I_D. We compose FADNet with two subnetworks, i.e., noise removal and dynamic filter generation. The noise removal subnetwork fol-

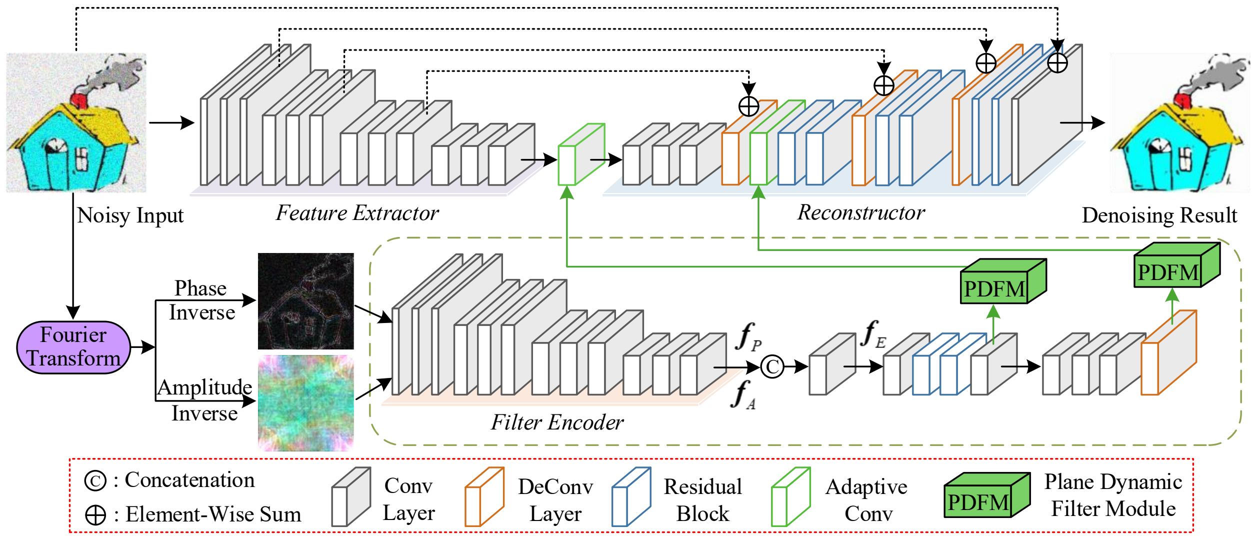


Figure 2: Overall architecture of the proposed FADNet.

lows an encoder-decoder architecture design. Given a noisy image $I_N \in \mathrm{R}^{H\times W\times 3}$, the feature extractor obtains the latent clean image feature $f_d \in \mathrm{R}^{h_d\times w_d\times c_d}$, where $w_d = \frac{W}{8}$ and $h_d = \frac{H}{8}$, and passes it through the reconstructor to restore a clean image. The dynamic filter generation subnetwork (DFGS) involves kernel construction, in which the noisy input is adaptively transformed to the weight kernels for feature modulation and update. Precisely, we first apply the Fourier transform algorithm (Frigo and Johnson 1998) on I_N and then perform the inverse Fourier transform to obtain the corresponding amplitude and phase images. The filter encoder then separately encodes the two inverse images into f_A and f_P, which are concatenated together, followed by a 1×1 convolution to form a feature map f_E. f_E is encoded further and finally fed into the plane dynamic filter module (PDFM). The PDFM is designed to predict the weight kernels. These predictions are convolved on the image's features, before being outputed to update the counterparts.

We next explain the network architecture. As shown in Fig. 2, there are mainly four types of layers (identified with different colors), viz., convolutional layer (Conv layer), deconvolutional layer (DeConv layer), residual block and adaptive Conv. More specifically, in the noise removal subnetwork, the Conv layer and rectified linear unit (ReLU) are adopted as the extractor to extract deep features from the input. While the Conv layer, adaptive Conv, DeConv layer and the residual block are the building mainstay of the reconstructor. Note that the residual block presented in (Ma et al. 2021b) is adopted in our FADNet. We also employ the symmetric skip connections to allow more abundant contextual features from the low level to be bypassed. As for the DFGS, we incorporate the Fourier transform, Conv layer, residual block and DeConv layer, in addition to the PDFM, for adaptive weight prediction.

In our work, we leverage the $L-2$ norm regularization to obtain the optimal parameters for the proposed FADNet. Given a training set of noisy-clean image pairs $\left\{I_N^j, I_G^j\right\}_{j=1}^{M}$, where I_N^j and I_G^j represent the j-th noisy and clean images, we optimize the proposed model by minimizing the following loss function:

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{M}\sum_{j=1}^{M}\left\|H_{FADNet}(I_N^j;\mathbf{\Theta}) - I_G^j\right\|_2^2, \qquad (1)$$

where $\mathbf{\Theta}$ denotes the parameters in training the denoising network and H_{FADNet} is the noisy-to-clean mapping function.

Dynamic Filter Generation Subnetwork (DFGS)

In this work, we aim to propose a well-generalized denoiser that works universally for the unknown and complicated real noises. The statistics of noise on real photographs, however, are usually spatially variant and chromatically correlated. Meanwhile, they are associated with different camera sensors and in-camera pipelines, which further scales up the complexity of real noise. Thus, using the pure encoder-decoder architecture may be inflexible for the complex noise and can be easily overfitted to a specific digital imaging device with certain noisy types. For this, we propose the DFGS to solve this dilemma. Our DFGS plays two crucial roles for the noise removal subnetwork. One is impeding the noise removal subnetwork from overfitting to limited training data, and the other is adapting the proposed denoiser to distinct noises in test noisy images dynamically and efficiently.

As depicted in Fig. 2, the proposed DFGS contains two main parts: Fourier transform and plane dynamic filter module.

Fourier transform For a single channel image I_N^S, its Fourier transformation can be expressed as:

$$\mathcal{F}(I_N^S)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} I_N^S(h,w)e^{-i2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \qquad (2)$$

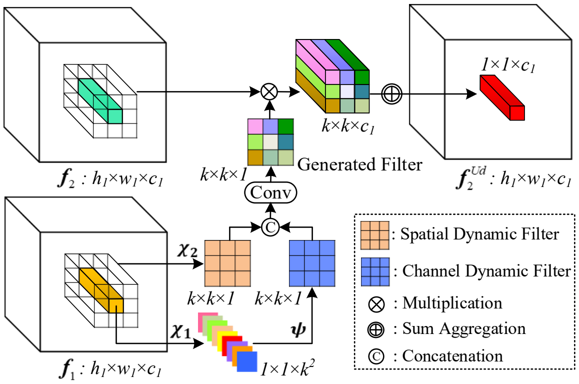


Figure 3: Schematic illustration of our proposed plane dynamic filter module. We set the convolution window as 3×3 in this example for easy demonstration.

where $i^2 = -1$. Let $\mathcal{A}(I_N^S)$ and $\mathcal{P}(I_N^S)$ be the amplitude and phase components of $\mathcal{F}(I_N^S)$, we have:

$$\begin{aligned} \mathcal{A}(I_N^S)(u,v) &= \left[R^2(I_N^S)(u,v) + C^2(I_N^S)(u,v)\right], \\ \mathcal{P}(I_N^S)(u,v) &= arctan\left[\frac{C(I_N^S)(u,v)}{R(I_N^S)(u,v)}\right], \end{aligned} \tag{3}$$

where $R(I_N^S)$ and $C(I_N^S)$ stand for the real and imaginary part of $\mathcal{F}(I_N^S)$, respectively. We then independently inverse $\mathcal{A}(I_N^S)$ and $\mathcal{P}(I_N^S)$ and map them back to their image space. For an RGB image, one can obtain the corresponding amplitude and phase inversed images by computing the Fourier transform for each channel individually.

As shown in Fig. 2, the $\mathcal{P}(I_N^S)$-only reconstruction produces the important semantic information and visual structures. While for the $\mathcal{A}(I_N^S)$-only reconstruction, it reveals the color-wise information conveyed in the original image. With simple network training, the filter encoder is encouraged to attain the essential semantic and color deep features without sophisticated network design. Finally, we integrate the output features to enhance the contextual aggregation for adaptive weight kernel prediction. Since the primary task of image denoising is to effectively eliminate the corrupted noise while reserving the desired fine image details, the semantic and color cues are of significant importance to facilitate the generated filters to adapt to the image contents with textural and structural information.

Plane dynamic filter module (PDFM) The PDFM has the purpose of enriching the convolution kernels with the ability to generalize well on diverse noisy input. As shown in Fig. 3, the PDFM is built upon a simple convolutional module, whose input is the feature maps, while the output is the weight kernels. In this subsection, we begin by introducing the standard dynamic filtering operation to draw forth our proposed PDFM.

The standard dynamic filters, predicted by separate network branches, are spatially varying at each pixel and capable of facilitating adaptive learning on image contents and feature embeddings. In general, a dynamic filter generation network takes a feature tensor f_{in} (with a size of $h_{in} \times w_{in}$ having c_{in} channels) as input and generates the filtered result $f^* \in \mathrm{R}^{k_h \times k_w \times c_{in} \times c_{out}}$, where c_{in} and c_{out} enumerate the input and output channels, respectively. The generated weight filters are then applied on the target image features in a sliding-window manner, and consequently, enable an adjustment of a deep network for flexible operators on-the-fly. However, predicting such a weight kernel often requires a large number of parameters (i.e., $k_h \times k_w \times c_{in} \times c_{out}$), which entails heavy side-networks and leads to huge memory consumption and computational loads. To achieve a well-generalized and efficient denoiser, the generated weight kernels are expected to be content-adaptive and lightweight. We accomplish this by proposing the PDFM, where the key idea behind our framework is to incorporate both the spatial and channel attention into the dynamic filter generation step.

As visualized in Fig. 3, let $f_1 \in \mathrm{R}^{h_1 \times w_1 \times c_1}$ be the input feature map of the PDFM, we start by generating both the channel and spatial dynamic filters from f_1. Precisely, for a single pixel $H \in \mathrm{R}^{1 \times 1 \times c_1}$ at spatial location (m, n) centered within a $k \times k$ convolution window, to obtain the channel dynamic filter $q \in \mathrm{R}^{k \times k \times 1}$, we have:

$$q = \boldsymbol{\psi}(\boldsymbol{\mathcal{X}}_1(H)). \tag{4}$$

In this formula, $\boldsymbol{\mathcal{X}}_1$ includes two operations: first it performs the global average pooling (GAP) across spatial dimensions and thus yields an aggregated contextual feature $\hat{q} \in \mathrm{R}^{1 \times 1 \times c_1}$. Then, we employ an 1×1 convolution layer to rescale the feature dimension from $1 \times 1 \times c_1$ to $1 \times 1 \times k^2$. $\boldsymbol{\psi}$ signifies the linear transformation that implements the channel-to-space rearrangement. As for the spatial dynamic filter $p \in \mathrm{R}^{k \times k \times 1}$, we apply two parallel operations, viz., the GAP and global max pooling (GMP), on the pixels along the channel dimensions within the $k \times k$ spatial neighborhood and finally aggregate the outputs from these two operations. We symbolize this generation procedure as $\boldsymbol{\mathcal{X}}_2$, which takes the following form:

$$p = \boldsymbol{\mathcal{X}}_2(H). \tag{5}$$

Following this, we concatenate the generated spatial and dynamic filters to form a filter $V \in \mathrm{R}^{k \times k \times 1}$ for better representation interweavement. As the generated filter V is conditioned on f_1, the exceedingly large or small filter values can be problematic in stabilizing the training of the network model. We followed (Zhou et al. 2021) to tackle this issue, using the filter normalization (FN) such that the filter values can be tweaked to a reasonable range.

We now turn to the step of applying the generated dynamic filter on the feature map of our reconstructor. Note that this procedure corresponds to the operations of the adaptive Conv as shown in Fig. 2. Concretely speaking, given a decoded feature map f_2 whose spatial size and channel number are the same as f_1, V is specially tailored for the pixel O located at (m, n). In other words, each generated filter is ingested into a single pixel for incarnation. More correctly, in the $k \times k$ spatial vicinity of the center pixel O, V executes the multiplication broadcast operation along all the channel dimensions, thus forming $\mathbf{E} \in \mathrm{R}^{k \times k \times c_1}$. To do so, the information in the channel dimension of a pixel lattice is explicitly associated with its neighborhood, after which the features in an enlarged receptive field can be well aggregated. The final new output is derived by a summation gathered operation over the spatial domain.

Our PDFM enjoys three desirable properties. First, the generated filter is conditioned on the input noisy image while being guided by the Fourier transform. Such a mechanism allows the PDFM to flexibly adapt to the image contents. Second, we tactfully decouple the filter generation process into its spatial and channel. This way, the spatial and channel information can be interwoven, collaboratively enhancing the representation capability. Third, the proposed PDFM offers a satisfactory solution to reduce the prohibitive memory and computation consumption, in which the parameter count of our predicted weight kernel is $k \times k$ and is far fewer than the commonly used dynamic filter generation based works. As we shall soon see in the experimental part, the FADNet inherites incredible advantages from PDFM by offering a considerably better denoising performance and generalization ability.

Experiments

Dataset

In this work, the training data was from SIDD (Abdelhamed, Lin, and Brown 2018). SIDD contained 30k high-resolution real noisy images captured by 5 smartphone cameras. The corresponding clean images were obtained by a systematic procedure. SIDD also provided a medium version package, in which 320 images pairs were leveraged for fast training and 1280 images pairs for validation purposes. In the experimental studies, we used 320 noisy-clean image pairs from the medium version of SIDD for network training.

At the test stage, in addition to SIDD, we further adopted three datasets to validate the effectiveness of FADNet. In particular, Nam (Nam et al. 2016) included noisy images of 11 static scenes, each of which was shot 500 times using the same consumer camera. By averaging these 500 shots from the same scene, the approximately noise-free images can be obtained. Considering that the images were of megapixel-size, we randomly cropped 300 smaller images (with a size of 512×512) for different scenes to perform the experiments. Darmstadt Noise Dataset (DND) (Plötz and Roth 2017) was composed of 50 real noisy images. All the images were collected under the outdoor lighting environment by 4 consumer-grade cameras. While the ground-truth clean data has not been released, one can submit the denoising results to the official online server and obtain the quantitative scores. We further introduced a new dataset, namely urban nightscape (UrbanN). UrbanN consisted of 50 noisy images that were acquired by three popular smartphones (i.e., iPhone 7 Plus, iPhone 12, and Huawei Mate12) under different urban landscapes at night. Since the noisy images in UrbanN had no ground truth, the visual comparisons were the main metric.

Implementation Details

We utilized the Adam optimizer (Kingma and Ba 2014) to update the network, with $\beta_1 = 0.7$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate was initially set as 0.001 and reduced to 0.0001 when the training errors held steady. In the dynamic filter generation subnetwork, we employed 2 PDFMs for feature regularization and found that using more PDFMs led to a slightly better performance, but entailed more computational loads at runtime. All the experiments were carried out using the Pytorch library (Paszke et al. 2019) on a machine with an NVIDIA Titan Xp GPU.

DFGS	–	✓	✓	✓	✓
Fourier	–	–	✓	✓	✓
Spatial	–	✓	–	✓	✓
Channel	–	✓	✓	–	✓
PSNR	35.29	39.08	40.19	40.06	**41.51**
SSIM	0.874	0.972	0.980	0.978	**0.992**

Table 1: Ablation study of four variations (i.e., DFGS, Fourier transform, spatial dynamic filter and channel dynamic filter) of our FADNet on the Nam dataset.

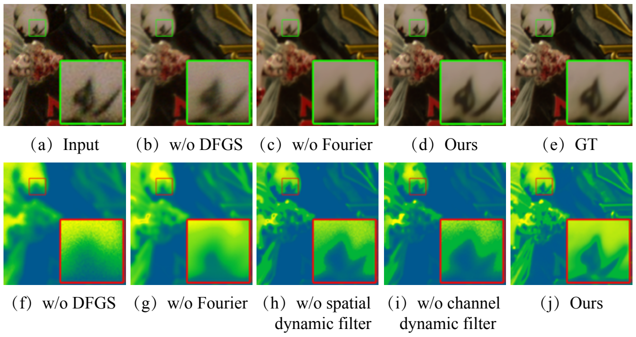

(a) Input (b) w/o DFGS (c) w/o Fourier (d) Ours (e) GT
(f) w/o DFGS (g) w/o Fourier (h) w/o spatial dynamic filter (i) w/o channel dynamic filter (j) Ours

Figure 4: Visualization (first row: denoising results; second row: heat maps) of the ablation study. It is worth noting that the heat maps interpret the representative capability of the filters generated by the second PDFM for the test image. Please view in color with zoom.

Ablation Study

In this subsection, several ablation studies were designed to investigate the impact of the following components: DFGS, the Fourier transform, and the generated filters (i.e., spatial and channel). Results are reported in Table 1 and Fig. 4, respectively.

Efficacy of the DFGS First of all, we validated the utility of the DFGS. Table 1 shows that excluding the DFGS from the proposed FADNet caused the largest performance drop, viz., from 41.51 dB to 35.29 dB. Furthermore, as shown in Fig. 4 (b), we note that without the DFGS, the network found it difficult to eliminate the real noise. Considering the complexity of the real noise, training the encoder-decoder architecture solely was not effective to fully exploit and learn the image characteristics. Conversely, the proposed FADNet by featuring the DFGS yielded noticeable improvements in visual quality, and consequently achieved higher PSNR and SSIM values, thereby supporting our claim about the advantage of using the DFGS.

Method	Nam PSNR / SSIM	SIDD PSNR / SSIM	DND PSNR / SSIM	Flops	Network Parameters	Running Speed
DnCNN-B (Zhang et al. 2017)	37.69 / 0.952	38.56 / 0.910	37.90 / 0.943	146.94	0.56	0.068
FFDNet+ (Zhang, Zuo, and Zhang 2018)	38.81 / 0.957	38.60 / 0.909	37.61 / 0.942	**107.29**	**0.48**	**0.030**
CBDNet (Guo et al. 2019)	39.08 / 0.969	38.68 / 0.909	38.06 / 0.942	144.36	4.34	0.422
RIDNet (Saeed and Nick 2019)	39.20 / 0.972	38.71 / 0.913	39.23 / 0.953	392.53	1.49	0.237
VDN (Yue et al. 2019)	39.68 / 0.976	39.28 / 0.909	39.38 / 0.952	158.49	7.81	0.519
DANet (Yue et al. 2020)	40.15 / 0.978	39.30 / 0.916	39.58 / 0.955	129.27	9.15	0.483
CycleISP (Zamir et al. 2020)	40.09 / 0.978	39.34 / 0.916	39.56 / 0.956	736.81	2.8	0.236
AINDNet (Kim et al. 2020)	39.98 / 0.978	39.45 / 0.915	39.53 / 0.956	271.19	13.76	0.577
MIRNet (Syed et al. 2020)	40.27 / 0.979	39.53 / 0.917	39.88 / 0.956	2372.93	31.78	0.885
InvNet (Liu et al. 2021)	40.15 / 0.978	39.38 / 0.916	39.57 / 0.952	112.49	2.64	0.214
NBNet (Cheng et al. 2021)	40.30 / 0.980	39.62 / 0.919	39.89 / 0.955	316.76	13.3	0.368
FADNet (Ours)	**41.51 / 0.992**	**39.96 / 0.926**	**40.17 / 0.959**	111.35	2.59	0.197

Table 2: Quantitative comparisons (PSNR(in dB) / SSIM, network parameters (in M), floating-point operations (Flops) (in G) and running speed (in second)) of our FADNet against other competitive approaches. The Flops and running speed were computed by processing the testing images with a size of 512×512 on the Nam dataset.

Efficacy of the Fourier transform We next assessed the necessity of the Fourier transform. In this work, we injected the amplitude and phase information of the Fourier transform into the filter generation procedure to contextually enrich the predicted filters with structures and color information. As expected, utilizing the Fourier transform benefited in handling the complex real noise, which resulted in a 2.43 dB improvement on the Nam dataset. In Fig. 4(c), it is evident that the absence of the Fourier transform affected the image restoration quality, where the denoised result tended to lose some structural contents and fine textural details. Compared with the heat maps in Fig. 4(g) and Fig. 4(j), again, one can observe that the Fourier transform was helpful in improving the representations with clearer details and object boundaries.

Efficacy of the generated filters Finally, we probed the effect of the generated filters. Particularly, we conducted experiments with two settings by predicting either the spatial or channel filter. As visualized in Fig. 4 (h) and Fig. 4(i), employing the spatial or channel filter individually could not correctly respond to the semantic information of the heavily structured objects. Meanwhile, some noises were not removed completely. This was mainly ascribed to less of the network generalizability. When we integrated the two types of filters together, as shown in Fig. 4(j), the network can achieve a positive improvement based on its semantic understanding whilst well eliminating the noise. In short, these visualizations verified that the generated spatial and channel filters can jointly bring great benefit to (*i*) highlight the semantic representations of the image; (*ii*) boost the network generalization ability to the unknown real noise.

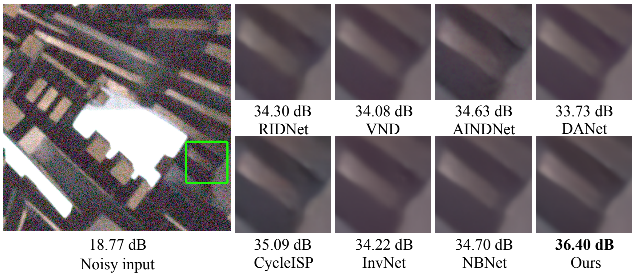


Figure 5: Denoised results of different algorithms on the noisy image from DND. Please view in color with zoom.

Main Results

Evaluation on Nam, SIDD and DND The average PSNR/SSIM results of all the competitive algorithms were listed in Table 2. From the analysis of the quantitative scores, it can be easily observed that: (*i*) The proposed FADNet noticeably advanced other competing methods by consistently achieving the best PSNR and SSIM values on three groups of testing data. Notice that in the PSNR metric, we obtained a performant rise over the second best method NBNet, boosting from 39.62 to 39.96 on the SIDD validation dataset, and from 39.89 to 40.17 on DND. (*ii*) Our FADNet excelled CBDNet, AINDNet and CycleISP by 0.5 dB at least, even though they embedded the noise information into their representation learning paradigms. (*iii*) The performance gains of our method over DnCNN-B and FFDNet+ became larger, possibly due to their overfitting issue to the Gaussian distribution.

We presented visualized comparisons on SIDD and DND in Figs. 1 and 5, from which we can see that the proposed FADNet was effective in removing the real noise and delivering visually pleasant and faithful results. In contrast, other approaches either compromised structural content, or produced blurring artifacts. In Fig. 1, though the input image was contaminated by large noise intensities, the denoised image by FADNet contained more apparent textural details than in other approaches. In Fig. 5, it can be seen that FAD-

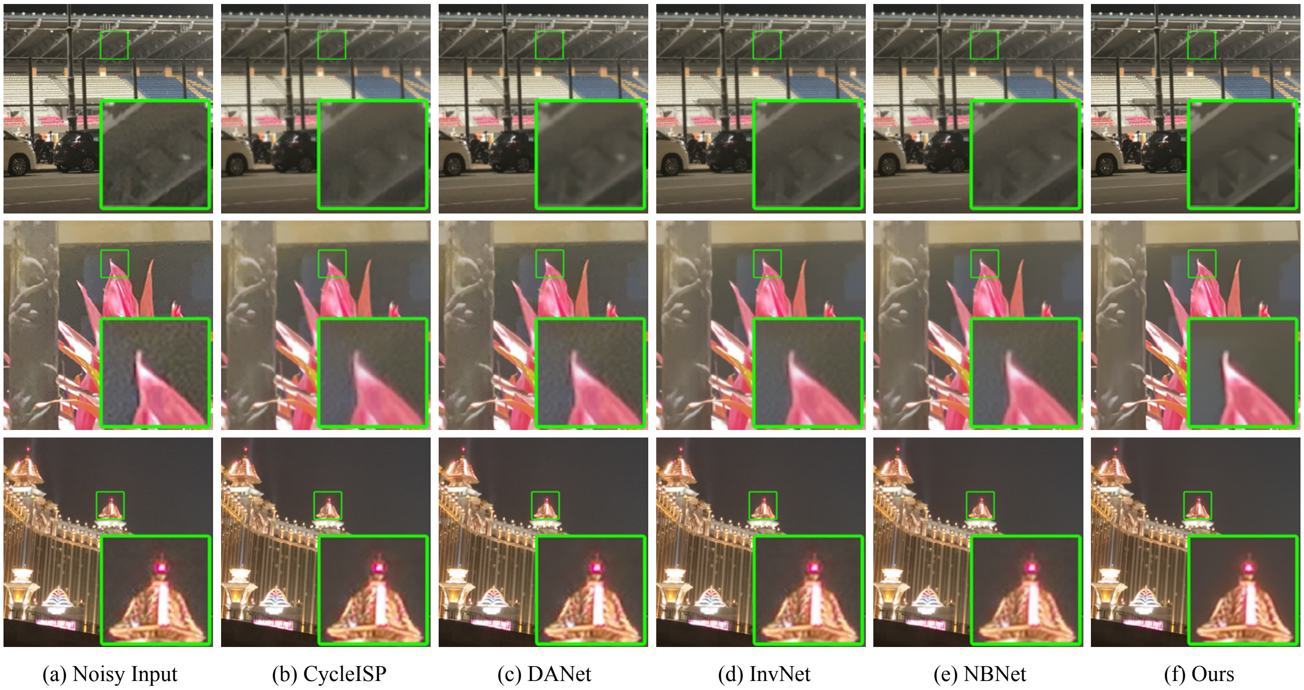

(a) Noisy Input (b) CycleISP (c) DANet (d) InvNet (e) NBNet (f) Ours

Figure 6: Comparison on UrbanN in the evaluation of generalization capabilities. The real noisy images, from top to bottom, were captured by iPhone 7Plus, iPhone 12 and Huawei Mate12, respectively. Despite the large noise discrepancy between the training and test data, our method consistently exhibited promising generalization abilities to the unknown and sophisticated real noise. Please view in color with zoom.

Net outperformed other algorithms in restoring more crisp edges and tiny details.

Evaluation on generalization capability To demonstrate the practicability, we next inspected the generalization capability of the proposed denoiser. It is noteworthy that the recent public benchmarks (i.e., Nam, DND and SIDD) and UrbanN were collected by different types of cameras. Based on the discussions in Section I, different camera sensors and ISP pipelines would give rise to the noise discrepancy. Moreover, since UrbanN did not contain any training images and ground-truth images, it was suitable for generalization testing.

We applied our FADNet on UrbanN with a comparison to the recent state-of-the-art methods. Fig. 6 illustrates the visual results. Notably, our algorithm can handle the unknown real noise flexibly, suppress the artifacts effectively and preserve the image details well. While for the other competing techniques, they failed to remove the noise since the noise domains of the training and test set did not coincide. Furthermore, the first row depicts that FADNet can better recover the subtle structures and textures, whereas the other methods suffered from serious blurring artifacts. Overall, compared with the state-of-the-art on UrbanN, the superiority of FADNet confirmed the better generalization capability to images with different noise domains.

Evaluation on efficiency In addition to the denoising performance, we also compared the network efficiency of the competing methods. All experiments were conducted on images with a size of 512×512 from the Nam dataset with the same GPU. As can be seen from Table 2, FFDNet+ had the fewset network parameters and Flops. Nevertheless, it obtained a limited performance. The network efficiency of our FADNet was more comparable with those of the recent competitive algorithms. In particular, compared to the most recently proposed NBNet, we provided much more satisfactory results, i.e., 41.5 dB vs 40.30 dB, by only taking 35.15% of its computational cost and 19.47% of its number of parameters, demonstrating appealing efficiency and effectiveness in real-world applications.

Conclusion

In this paper, we proposed a novel denoiser, termed as FADNet, for flexible and adaptive real-world noisy image denoising. To realize this goal, we applied the encoder-decoder architecture, Fourier transform, and dynamic filter techniques into in the network design. The results from the ablation study validated the necessity and effect of the key components of our network model. The results on Nam, SIDD, DND and UrbanN showed that the proposed FADNet can not only cope with the sophisticated real noise, but also exhibit promising generalization abilities to images with noise discrepancy, achieving superiority over the state-of-the-art methods. The efficiency comparisons further demonstrated the need for fewer parameters and lower computational cost of FADNet against the recent best algorithms. Considering its robustness, adaptation and efficiency, FADNet is essentially more feasible for practical denoising applications.

Acknowledgments

This work was supported in part by the University of Macau (File No. MYRG2019-00006-FST), the Youth Innovation Project of the Department of Education of Guangdong Province, China (File No. 2020KQNCX040), the Special Projects for Key Fields in Higher Education of Guangdong, China (File No. 2021ZDZX1042).

References

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A High-Quality Denoising Dataset for Smartphone Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1692–1700.

Chandran, P.; Zoss, G.; Gotardo, P.; Gross, M.; and Bradley, D. 2021. Adaptive Convolutions for Structure-Aware Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7972–7981.

Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; and Liu, S. 2021. NBNet: Noise Basis Learning for Image Denoising With Subspace Projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4896–4906.

Frigo, M.; and Johnson, S. G. 1998. FFTW: an adaptive software architecture for the FFT. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 3, 1381–1384.

Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; and Zhang, L. 2019. Toward Convolutional Blind Denoising of Real Photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1712–1722.

Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; and Sun, J. 2019. Meta-SR: A Magnification-Arbitrary Network for Super-Resolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1575–1584.

Jia, X.; Liu, S.; Feng, X.; and Zhang, L. 2019. FOCNet: A Fractional Optimal Control Network for Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6047–6056.

Karaimer, H. C.; and Brown, M. S. 2016. A software platform for manipulating the camera imaging pipeline. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 429–444.

Kim, S. J.; Lin, H. T.; Lu, Z.; Ssstrunk, S.; Lin, S.; and Brown, M. S. 2012. A New In-Camera Imaging Model for Color Computer Vision and Its Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12): 2289–2302.

Kim, Y.; Soh, J. W.; Park, G. Y.; and Cho, N. I. 2020. Transfer Learning From Synthetic to Real-Noise Denoising With Adaptive Instance Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3479–3489.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lee, J.; Son, H.; Rim, J.; Cho, S.; and Lee, S. 2021. Iterative Filter Adaptive Network for Single Image Defocus Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2034–2042.

Liu, Y.; Qin, Z.; Anwar, S.; Ji, P.; Kim, D.; Caldwell, S.; and Gedeon, T. 2021. Invertible Denoising Network: A Light Solution for Real Noise Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13365–13374.

Ma, R.; Hu, H.; Xing, S.; and Li, Z. 2020. Efficient and Fast Real-World Noisy Image Denoising by Combining Pyramid Neural Network and Two-Pathway Unscented Kalman Filter. *IEEE Transactions on Image Processing*, 29: 3927–3940.

Ma, R.; Li, S.; Zhang, B.; and Li, Z. 2021a. Towards Fast and Robust Real Image Denoising with Attentive Neural Network and PID Controller. *IEEE Transactions on Multimedia*, 1–13.

Ma, R.; Zhang, B.; Zhou, Y.; Li, Z.; and Lei, F. 2021b. PID Controller-Guided Attention Neural Network Learning for Fast and Effective Real Photographs Denoising. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Nam, S.; Hwang, Y.; Matsushita, Y.; and Kim, S. J. 2016. A Holistic Approach to Cross-Channel Image Noise Modeling and Its Application to Image Denoising. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1683–1691.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 8026–8037.

Plötz, T.; and Roth, S. 2017. Benchmarking Denoising Algorithms with Real Photographs. In *Proceedings of IEEE/VCF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2750–2759.

Saeed, A.; and Nick, B. 2019. Real Image Denoising With Feature Attention. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3155–3164.

Seybold, T.; Cakmak, .; Keimel, C.; and Stechele, W. 2014. Noise characteristics of a single sensor camera in digital color image processing. *Color and Imaging Conference*, 6: 33–58.

Syed, Z., Waqas; Aditya, A.; Salman, K.; Munawar, H.; Fahad, K., Shahbaz; Yang, M.-H.; and Yang, M.-H. 2020. Learning Enriched Features for Real Image Restoration and Enhancement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 492–511.

Tsin, Y.; Ramesh, V.; and Kanade, T. 2001. Statistical calibration of CDD imaging process. In *Proceedings of IEEE*

International Conference on Computer Vision (ICCV), volume 1, 480–487.

Xu, C.; Wu, B.; Wang, Z.; Zhan, W.; Vajda, P.; Keutzer, K.; and Tomizuka, M. 2020. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, 1–19.

Xu, L.; Zhang, J.; Cheng, X.; Zhang, F.; Wei, X.; and Ren, J. 2021. Efficient Deep Image Denoising via Class Specific Convolution. In *Proceedings of the Third-Fifth AAAI Conference on Artifical Intelligence*, 3039–3026.

Yue, Z.; Yong, H.; Zhao, Q.; Zhang, L.; Ozair, S.; and Meng, D. 2019. Variational denoising network: Toward blind noise modeling and removal. In *Proceedings of the IEEE Conference on Neural Information Processing Systems (NIPS)*, 1690–1701.

Yue, Z.; Zhao, Q.; Zhang, L.; and Meng, D. 2020. Dual Adversarial Network: Toward Real-world Noise Removal and Noise Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 41–58.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2020. CycleISP: Real Image Restoration via Improved Data Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2693–2702.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7): 3142–3155.

Zhang, K.; Zuo, W.; and Zhang, L. 2018. FFDNet: Toward a Fast and Flexible Solution for CNN-Based Image Denoising. *IEEE Transactions on Image Processing*, 27(9): 4608–4622.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2021. Residual Dense Network for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2480–2495.

Zhou, J.; Jampani, V.; Pi, Z.; Liu, Q.; and Yang, M.-H. 2021. Decoupled Dynamic Filter Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6647–6656.