Towards Robust Image Denoising via Flow-based Joint Image and Noise Model

Lanqing Guo, Siyu Huang, *Member, IEEE,* Haosen Liu, and Bihan Wen, *Member, IEEE*

***Abstract*—One of the fundamental challenges in image restoration is denoising, where the objective is to estimate the clean image from its noisy measurements. Existing denoising approaches generally focus on exploiting effective natural image priors to remove the noise. However, the utilization and analysis of the noise model are often ignored, although the noise model can provide complementary information to the denoising algorithms. As a result, they are very sensitive to different noise distributions. To tackle this issue and hence towards a robust image denoiser in practice, in this paper, we propose a novel Flow-based joint Image and NOise model (FINO) that distinctly decouples the image and noise in the latent space and losslessly reconstructs them via a series of invertible transformations. We further present a variable swapping strategy to align structural information in images and a noise correlation matrix to constrain the noise based on spatially minimized correlation information. Experimental results demonstrate FINO's capacity to remove both synthetic additive white Gaussian noise (AWGN) and real noise. Furthermore, the generalization of FINO to the removal of spatially variant noise and noise with inaccurate estimation surpasses that of the popular and state-of-the-art methods by large margins.**

***Index Terms*—Image Denoising, Invertible Neural Network, Real Noise, Synthetic Noise, Disentanglement**

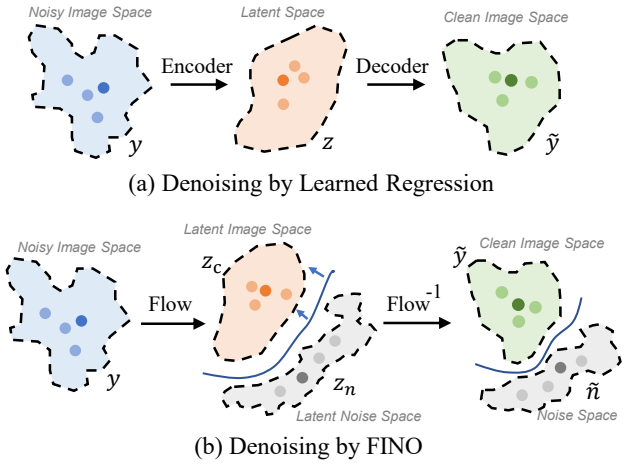


Fig. 1: The framework of Flow-based joint Image and NOise model (FINO) and comparison between previous regression-based denoisers (a) and the proposed FINO (b).

I. Introduction

IMAGE denoising refers to recovering the underlying clean image from an observed noisy measurement. Despite today's vast improvement in camera sensors, digital images are often corrupted by severe noises in complex environments, resulting in nontrivial effects to subsequent vision tasks. Thus, image denoising is a crucial task that may significantly affect the subsequent vision tasks.

Existing image denoising methods generally rely on the construction of effective image priors. For conventional methods, the corresponding priors include, *e.g.,* sparsity [1], [2], low-rankness [3]–[5], and non-local similarity [6], [7]. By shrinkage or filtering in the transform domain, image components that are satisfied with the prior are preserved in the denoised results. However, the non-learning based transform is limited, *e.g.,* discrete cosine transform (DCT), and the learning-based transform methods are more flexible but they all worked on patches, lacking of global modeling. Recently, the deep learning approaches [8]–[11] have achieved state-of-the-art image denoising results by relying on an external training corpus. These deep denoisers generally learn a direct mapping from noisy images to clean images, where the learned models serve as an effective prior on the clean image space. Despite the success of learning effective image priors for denoising, few work to-date investigated the noise modeling that is complementary to the deep image priors learning. In practice, the residual between the noisy image and the denoised one always contains image structures that are wrongly removed together with noise. These structures generally correspond to high-frequency components of an image, which are distinct from the random noise. This inspires us that it might be possible to constrain the residual map with the noise model to 'squeezing' image information from the residual map. Such stricter modeling on both image and noise can also decrease the overfitting to training corpus and improve the robustness of denoisers.

Different from all existing deep denoising methods, we reformulate the image denoising task as a dual modeling problem of both the image and noise (as illustrated in Figure 1(b)), instead of only reconstructing the noise-free image (Figure 1(a)). However, *how to achieve an effective decoupling of clean image and noise* is a challenging and under-explored problem in the existing literature. In this paper, we show that the noisy images can be losslessly transformed to a more distinguishable feature space through a novelly proposed framework named **F**low-based joint **I**mage and **NO**ise model (**FINO**). FINO losslessly decouples the image-noise compo-

L. Guo, and B. Wen are with School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore 639798. E-mail: lanqing001@e.ntu.edu.sg, bihan.wen@ntu.edu.sg.

S. Huang is with the Visual Computing Division, School of Computing, Clemson University, Clemson, SC 29631 USA. E-mail: siyuh@clemson.edu.

H. Liu is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong. E-mail: hsliu@eee.hku.hk.

nents in the latent space through a forward process of the flow-based invertible network. Then, the decoupled components can be reconstructed as the noise and image in the spatial domain through a backward process of an invertible network. Based on the decoupled noise and image components, we further introduce a noise variable swapping strategy to align the structural information in images, as well as a constraint on the noise correlation matrix to be spatially independent on the neighboring regions. Extensive experiments are conducted to evaluate FINO on both the synthetic noise and real noise removal tasks. Empirical results show that FINO achieves superior performances in comparison with the state-of-the-art image denoising methods. Besides, FINO provides significantly better generalizability and robustness than the existing denoising methods.

The contributions of this work are summarized as follow:

- We propose to jointly model the distributions of the image and noise for denoising tasks, showing that the noise model can provide abundant and complementary information, in addition to image priors.
- We present a novel image denoising framework named FINO which distinctly disentangles the noise and the noise-free image in the latent space. Two learning methods, including variable swapping and noise correlation matrix, are also proposed to improve the learning of FINO.
- We conduct extensive experiments on both the synthetic and real noise datasets. FINO shows superior denoising and generalization performances compared to the existing denoising methods.

The rest of the paper is organized as follows: Section II introduces related image denoising methods, neural flow models, and disentangled feature representations for computer vision. In Section III, the preliminary knowledge of the multi-scale neural flow is introduced. In Section IV, the problem formulation of robust image denoising is introduced, the discussion of why employing the invertible neural networks for image-noise disentangling is presented, and the architecture details of proposed FINO is described. Experimental results are shown in Section V and concluding remarks are given in Section VI.

II. Related Work

A. Image Denoising

Model-based image denoising. Image denoising is a typical ill-posed problem with the goal of recovering high-quality images from their noisy measurements. Numerous efforts have been made towards it over the past decades. Classic methods generally take advantage of the image priors, such as sparsity [1], [2], low rank [3], and non-local self-similarity [12]–[14], to address the denoising problem. Most classic denoisers utilize the image features in certain transform domains by applying shrinkage or filtering to the exploitation of image priors. The learning-based transform methods are more flexible, but they all worked on patches, lacking global modeling. For instance, BM3D [15] applies effective filtering in 3D transform domain by combining sliding-window transform processing with block matching. WNNM [3] incorporates low-rank matrix approximations using the weighted nuclear norm. Liu *et al.* [16] proposed the CAS algorithm to exploit the local and non-local similar blocks by a group of similar patches that are extracted from clustered rows of patch groups that consist of similar image contents. Motivated by this, in this work, we employ a flow-based invertible network to conduct a learnable global transforming.

Apart from the image priors, traditional denoisers also make assumptions on noise, where they formulate MAP estimations when noise distribution is known, *e.g.,* ℓ_1 and ℓ_2 norms for Laplacian and Gaussian noise, respectively. Recent works exploited hybrid noise similarly, *e.g.,* Meng *et al.* [17] and Cao *et al.* [18] assumed noise as mixture of exponential power distributions in optimization.These models are either too specific or hard to optimize. In contrast, FINO explicitly models noise representation in **deep learning** with only **mild** assumption, *i.e.,* independent noise, for the first time.

Deep learning-based image denoising. In recent years, deep learning-based denoisers exhibit superiority in learning the end-to-end mapping from noisy to clean images [8]–[10], [19], [20]. For instance, Schuler *et al.* [21] employed multi-layer perceptron (MLP) for image denoising and achieves better denoising performance than classical BM3D method [15]. After that, Zhang *et al.* [8] proposed a convolutional neural network based denoiser, which achieves a very competitive denoising performance through residual neural networks. Zhang *et al.* [9] further introduces a noise level map to control the trade-off between noise reduction and detail preservation. To exploit the non-local property of the image features in deep convolutional neural network, Plotz *et al.* [22] presents an N3Net by employing the k-nearest neighbor matching in the denoising network. After that, transformer-based models [23] take advantage of the long-range dependencies within the context, which also have gained improvements among various vision tasks. Some researchers [24]–[26] try to leverage the transformer based architectures to image restoration, achieving superior performance, while those methods are always time-consuming with large parameters. Generally, existing denoising methods focus on exploiting effective natural image priors, while the modeling, analysis, and utilization of the noise component are often ignored. As a result, those denoisers are very sensitive to different noise distributions. This work jointly models the natural image and noise via invertible neural networks to deliver better denoising performance and visual quality.

More recently, there are some attempts [27]–[30] for denoising on real noisy images. The attempts can be generally divided into two categories: 1) Two-step denoising [27], which first estimates the noise map then reconstructs the clean image non-blindly based on the estimated noise map. For instance, VDN [27] learns an approximate posterior to the true posterior with the latent variables, which mainly focuses on non-i.i.d. noise distribution; 2) One-step denoising with an end-to-end framework [28]. Typically, RIDNet [28] jointly learns the noise and denoiser without using a separate noise estimation branch. This work focuses on both real noise and synthetic noise removal. It follows the one-step denoising approaches

to enhance the generalization capacity of denoising models.

Comparison with InvDN [31]. Very recently, Liu *et al.* [31] propose to adopt the invertible network for image denoising, dubbed InvDN. The InvDN is based on separating low/high-frequency components and discarding the high-frequency entangled with noise. However, directly removing high-frequency components often results in the loss of important image structures that are challenging to reconstruct. In contrast to this approach, we propose to decouple and collect the noise and image components via the invertible network, allowing us to model these two components individually. In this way, the high-frequency component is well retained with our method.

B. Neural Flows

The neural flow is a type of deep generative model that learns the exact likelihood of targets through a chain of reversible transformations. The generative process $\mathbf{x} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{z})$ given a latent variable $\mathbf{z}$ can be specified by an invertible architecture $\mathcal{F}_{\boldsymbol{\theta}}$. The direct access to the inverse mapping is $\mathbf{z} = \mathcal{F}_{\boldsymbol{\theta}}^{-1}(\mathbf{x})$. As a pioneering work, NICE [32] learns a highly non-linear bijective transformation that maps the training data to a space where its distribution is factorized. Following NICE, more effective and flexible transformations have been proposed [33]–[35].

More recently, a series of works [31], [36]–[38] exploit neural flows for image restoration, which formulate image restoration as a non-degradation image generation problem. For instance, SRFlow [37] designs a conditional normalizing flow architecture for super-resolution, which learns the distribution of realistic HR images. [38] and [31] apply invertible networks to image rescaling and image denoising tasks, respectively. InvDN [31] focuses on real noise removal and detours the noise-image disentanglement. It splits noisy images into low-frequency and high-frequency components and then directly drops the high-frequency component, restoring the image based on the low-frequency component only, which may result in information loss and over-smoothness. Different from InvDN, our proposed FINO utilizes an invertible network to decouple the image content and noise components in the latent space and then reconstruct them in the image space, respectively, achieving a lossless image-noise disentanglement.

C. Disentangled Feature Representations

Disentangled feature representations aim at learning an interpretable representation for image variants, which has been widely studied in various tasks, *e.g.,* face editing [39], image restoration [40], image classification [41], [42], as well as image translation [43]–[45]. UNIT [43] makes a shared-latent space assumption based on coupled GANs. As follows, to improve the diversity of output, models such as MUNIT [44], DRIT [45] are proposed to embed images onto domain-invariant content space and domain-specific attribute space via disentanglement. Choi *et al.* [46] further proposed a StarGAN can perform image-to-image translations for multiple domains using only a single model. Similarly, Liu *et al.* [47] proposed a UFDN that learns domain-invariant representation from multiple domains and can perform continuous cross-domain image translation and manipulation. However, directly applying existing disentanglement to image denoising would significantly affect the image-noise decoupling. Once the input noisy image is embedded into latent space, some non-structural information, *e.g.,* noise, is hard to be preserved. Different from previous disentanglement network, we employ invertible neural network to build the shared encoder-decoder to avoid the information loss.

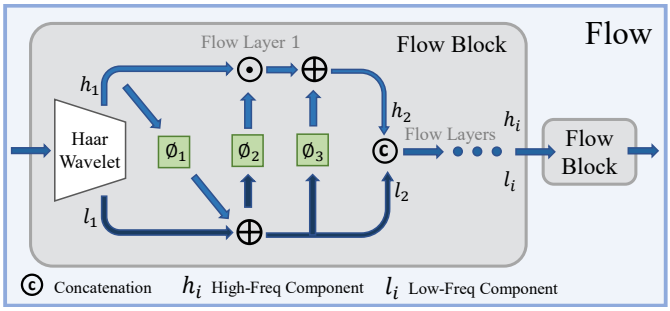


Fig. 2: The architecture of neural flow model, consisting of two flow blocks. Each flow block includes a invertible Haar wavelet transformation at first layer and twelve affine coupling layers. i denotes i-th affine coupling layer.

III. Preliminary

In this section, we introduce the preliminary knowledge of the multi-scale neural flow [31], [38], [48]. We denote it as $\mathrm{Flow}(\cdot)$ in this paper. As shown in Figure 2, $\mathrm{Flow}(\cdot)$ consists of a series of flow blocks, and each flow block consists of an invertible wavelet transformation followed by a series of affine coupling layers.

Invertible wavelet transform. To disentangle the information of clean image and noise, we employ invertible Haar wavelet transformation at the first layer of each flow block to downsample the input images/features and to increase the feature channels [38]. After the wavelet transformation, the input image/features with a shape of (H, W, C) should be squeezed into $(H/2, W/2, 4C)$. $4C$ denotes three directions of high-frequency coefficients and one low-frequency representation [49]. The invertible wavelet transformation provides the separated low and high-frequency information to the following invertible neural layers.

Affine coupling layers. After the wavelet transformation layer, the input image/feature $\mathbf{u}_i$ has been splitted into low and high-frequency components, denoted as $\mathbf{h}_i$ and $\mathbf{l}_i$, respectively. We leverage the coupling layer [32] to further decouple the structural information and the degradation bias. Suppose the i-th coupling layer's input is $\mathbf{u}_i$ and the output is $\mathbf{u}_{i+1}$ $(i = 1 \dots I)$, the forward procedure in this block is

$$\begin{aligned} \mathbf{l}_i, \mathbf{h}_i &= \mathrm{Split}(\mathbf{u}_i) \,, \\ \mathbf{l}_{i+1} &= \mathbf{l}_i + \phi_1(\mathbf{h}_i) \,, \\ \mathbf{h}_{i+1} &= \phi_2(\mathbf{l}_{i+1}) \odot \mathbf{h}_i + \phi_3(\mathbf{l}_{i+1}) \,, \\ \mathbf{u}_{i+1} &= \mathrm{Concat}(\mathbf{l}_{i+1}, \mathbf{h}_{i+1}) \,, \end{aligned} \tag{1}$$

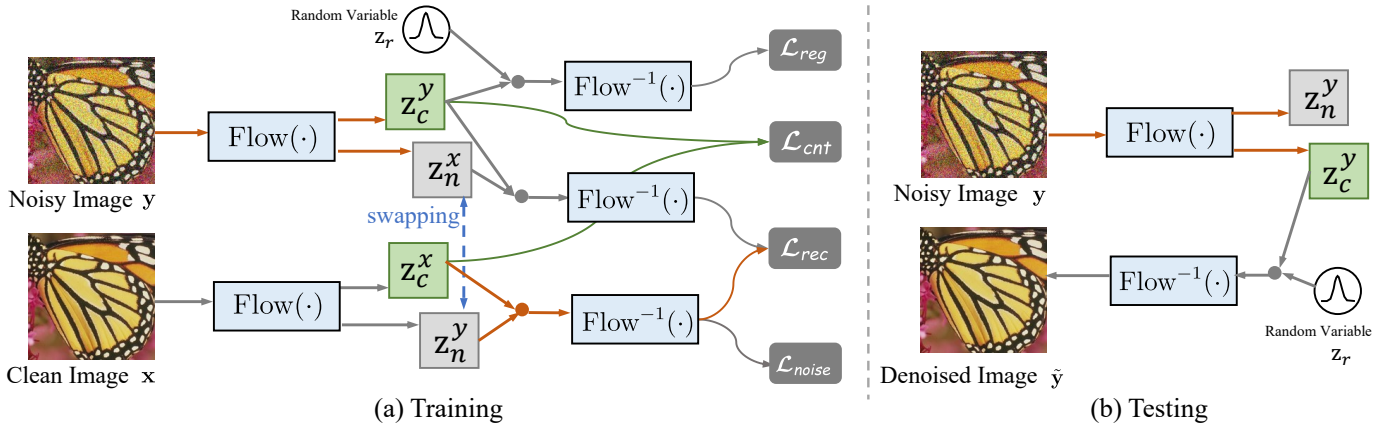


Fig. 3: An overview of our Flow-based joint Image and NOise model (FINO). In the training stage (a), the noisy and clean image pair are fed into the Flow(·), and we obtain the decoupled image and noise variables, *i.e.,* $\mathbf{z}_c^x$ and $\mathbf{z}_n^x$ from clean image and $\mathbf{z}_c^y$ and $\mathbf{z}_n^y$ from noisy image. A noise variable swapping strategy is employed to reconstruct the noise and image. We then introduce noise loss function to constrain the distribution of noise. In the testing stage, combining the disentangled image variable and sampled random variable from a known distribution and applying inverse flow model to generate the denoised image $\tilde{\mathbf{y}}$.

where $\mathrm{Split}(\cdot)$ denotes channel-wise splitting and $\mathrm{Concat}(\cdot)$ is the corresponding inverse operation. $\phi_1(\cdot)$, $\phi_2(\cdot)$, and $\phi_3(\cdot)$ can be any neural networks that are not required to be invertible. The backward procedure is easily derived as

$$\begin{aligned}\mathbf{l}_{i+1}, \mathbf{h}_{i+1} &= \mathrm{Split}(\mathbf{u}_{i+1}) \,, \\ \mathbf{h}_i &= (\mathbf{h}_i - \phi_3(\mathbf{l}_{i+1}))/\phi_2(\mathbf{l}_{i+1}) \,, \\ \mathbf{l}_i &= (\mathbf{l}_{i+1} - \phi_1(\mathbf{h}_i))/\phi_1(\mathbf{h}_i) \,, \\ \mathbf{u}_i &= \mathrm{Concat}(\mathbf{l}_i, \mathbf{h}_i) \,. \end{aligned} \tag{2}$$

IV. FINO

In this section, we present a neural flow-based framework named FINO to jointly model the image and noise in context of image denoising. We first formulate the image denoising problem, then discuss why employing the invertible neural networks for image-noise disentangling, and introduce the architecture details of FINO, including the variable swapping in latent space for disentanglement, the clean image regression for image modeling, and the noise correlation matrix for noise modeling, as well as objective functions.

A. Problem Formulation

A noisy image $\mathbf{y}$ and its noise-free counterpart $\mathbf{x}$ can be formulated as

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \,, \tag{3}$$

where $\mathbf{n}$ denotes the random noise. Here, we consider both the synthetic noise and the real noise. The synthetic noise is the *i.i.d* additive white Gaussian noise (AWGN). It follows the normal distribution $\mathcal{N}(0, \sigma^2 \cdot \mathbf{I})$. Typical supervised regression-based denoising methods train the deep neural networks (DNNs) by

$$\min_{\boldsymbol{\theta}} \ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathcal{L}_{reg}\left(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{x}\right)] \,, \tag{4}$$

where $\mathcal{G}_{\boldsymbol{\theta}}(\cdot)$ denotes the regression-based denoiser, which can be regarded as the combination of encoder and decoder $\mathcal{D}_{\boldsymbol{\theta}}(\mathcal{E}_{\boldsymbol{\theta}}(\cdot))$. $\mathcal{L}_{reg}(\cdot,\cdot)$ denotes the loss function, *e.g.,* the ℓ_1 or ℓ_2 loss.

Different from the conventional methods, we novelly propose a dual modeling of both noise and image as

$$\min_{\boldsymbol{\theta}} \ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathcal{L}_{reg}\left(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{x}\right) + \mathcal{L}_{noise}\left(\mathcal{H}_{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{n}\right)] \,, \tag{5}$$

where $\mathcal{L}_{noise}(\cdot,\cdot)$ denotes the loss function of the noise modeling, and $\mathcal{H}_{\boldsymbol{\theta}}(\cdot)$ denotes the noise model. The proposed FINO disentangles noise and image in a more distinguishable latent space via an invertible network, on behalf of $\mathcal{G}_{\boldsymbol{\theta}}$ and $\mathcal{H}_{\boldsymbol{\theta}}$, by jointly modeling image and noise.

B. Why Disentangling via Invertible Network?

Although it is known that the representation ability of the invertible network is limited and weaker than some more sophisticated deep neural networks, because of its specially designed structure [50], there are three main reasons why we choose the invertible network for image-noise decoupling.

- Firstly, existing regression-based image denoising algorithms cannot achieve lossless image reconstruction of the input image: Once the input noisy image is embedded into latent space, some information may get lost. This problem is especially severe for the non-structural information, *e.g.,* noise. Different from them, the invertible network can losslessly encode the image and noise. This property ensures that recoveries of the image and noise can always complement each other, providing the basis of improving the image denoising performance via a joint image and noise modeling.
- Secondly, only the forward module of invertible network needs to be trained, while the backward module is the direct inverse, which respectively act as encoder and decoder. The number of free parameters can be thus significantly reduced.
- Finally, image denoising is an ill-posed inverse problem of one-to-many mapping, *i.e.,* one noisy image can be

restored to many denoised estimations. Most of the existing methods formulate it as a one-to-one mapping task, *i.e.,* delivering one denoised image from one noisy input. However, FINO can sample diverse denoised images by coupling the disentangled clean image with any random variable sampled from a normal distribution.

C. The Framework of FINO

Given a noisy image $\mathbf{y}$ and its corresponding clean counterpart $\mathbf{x}$ in training stage, we first employ an invertible flow model $\mathrm{Flow}(\cdot)$ to embed the input pair to latent codes respectively, as

$$\mathbf{z}^x = \mathrm{Flow}(\mathbf{x}) \qquad \mathbf{z}^y = \mathrm{Flow}(\mathbf{y}) \ . \tag{6}$$

Variable swapping for disentangling. We divide the latent space $\mathcal{Z}$ into two sub-spaces, *i.e.,* clean image space and noise space, with separated latent codes $\mathbf{z} = [\mathbf{z}_c, \mathbf{z}_n]$. As shown in Figure 3(a), in order to ensure the noise information is decoupled from the noisy inputs, we introduce a noise variable swapping strategy to generate noisy image, *i.e.,* combining the noise variable $\mathbf{z}_n^y$ from noisy input and the clean image variable $\mathbf{z}_c^x$ from the clean counterpart as follows,

$$\hat{\mathbf{y}} = \mathrm{Flow}^{-1}(\mathbf{z}_c^x, \mathbf{z}_n^y) \qquad \hat{\mathbf{x}} = \mathrm{Flow}^{-1}(\mathbf{z}_c^y, \mathbf{z}_n^x) \ , \tag{7}$$

where $\hat{\mathbf{y}}$ is the reconstructed noisy image, $\hat{\mathbf{x}}$ is the reconstructed clean image, and $\mathrm{Flow}^{-1}(\cdot,\cdot)$ denotes the inverse process of $\mathrm{Flow}(\cdot)$. Following Equation (7), we can easily derive the reconstructed noise $\hat{\mathbf{n}} = \hat{\mathbf{y}} - \mathbf{x}$.

To ensure noise information is completely embedded in the noise variable $\mathbf{z}_n$, we impose a *reconstruction loss* to encourage recovery of the original noise $\mathbf{n}$ and clean image $\mathbf{x}$, using the noise and clean image variables, respectively:

$$\mathcal{L}_{rec} = \|\hat{\mathbf{n}} - \mathbf{n}\|_1 + \|\hat{\mathbf{x}} - \mathbf{x}\|_1 \ . \tag{8}$$

Besides, to enforce the clean image code $\mathbf{z}_c$ to contain only the noise-free content information, we employ a *content alignment loss* to align the structural features of noisy and clean image pairs, as

$$\mathcal{L}_{cnt} = \|\mathbf{z}_c^x - \mathbf{z}_c^y\|_1 \ . \tag{9}$$

Clean image regression. In addition, the denoised images can also be generated via its disentangled clean image component $\mathbf{z}_c^y$ and a random variable $\mathbf{z}_r$ as follow

$$\tilde{\mathbf{y}} = \mathrm{Flow}^{-1}(\mathbf{z}_c^y, \mathbf{z}_r) \ . \tag{10}$$

Note that the $\mathbf{z}_r$ would be a random variable sampled from a normal distribution or a zero variable in the reference stage. We employ a *regression loss* on the generated $\tilde{\mathbf{y}}$ and further enforce the noise variable to be independent of the structure information, as

$$\mathcal{L}_{reg} = \|\tilde{\mathbf{y}} - \mathbf{x}\|_1 \ . \tag{11}$$

Besides, the restored image can be sampled by combing their internal clean image variable and a normal distribution variable in the testing stage as shown in Figure 3(b).

Noise correlation matrix. We assume that the additive noise $\mathbf{n}$ is spatially uniform and uncorrelated (*e.g., i.i.d.* Gaussian noise). Let $V : \mathbf{n} \mapsto V\mathbf{n} \in \mathbb{R}^{m \times M}$ be an overlapping patch extractor, where m denotes the number of pixels within one patch and M denotes the number of patches. We obtain the noise patch matrix $\hat{\mathbf{N}} = V\hat{\mathbf{n}} = [\hat{\mathbf{N}}_1, \hat{\mathbf{N}}_2, \ldots, \hat{\mathbf{N}}_M]$ from the reconstructed $\hat{\mathbf{n}}$. The patch-wise noise correlation matrix is defined as

$$\mathbf{\Sigma} = \frac{1}{M} \sum_{j=1}^{M} \hat{\mathbf{N}}_j \hat{\mathbf{N}}_j^T \ . \tag{12}$$

Based on the uncorrelated noise assumption, all of the non-diagonal elements of $\mathbf{\Sigma}$ should be as close to zero as possible. Denote the standard deviation of $\mathbf{n}$ to be σ, the diagonal elements of $\mathbf{\Sigma}$ should all be σ^2. Therefore, we set the following *noise correlation loss* as

$$\mathcal{L}_{noise} = \|\mathbf{\Sigma} - \sigma^2 \mathbf{I}\|_F^2 \ . \tag{13}$$

Full objective function. By combining the above losses, the hybrid objective function $\mathcal{L}$ used to train our model is

$$\mathcal{L} = \underbrace{\mathcal{L}_{reg}}_{\text{image}} + \underbrace{\alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{cnt}}_{\text{disentangling}} + \underbrace{\gamma\mathcal{L}_{noise}}_{\text{noise}} \ , \tag{14}$$

where α, β, and γ are the weighting coefficients to balance the influence of each term.

D. Extension to Real Noise

Real noises are mostly spatially-correlated and spatially-variant, which is different from AWGN with pixel-independent. Inspired by [53], the domain gap between real noise and AWGN can be reduced by Pixel-shuffle Downsampling (PD) [54] strategy, where the spatially-correlated noises are broken down to pixel-wise independent noises. Thus, our uncorrelated noise assumption can be applied to the processed pixel-wise independent noises. We define the $\mathrm{PD}(\cdot)$ operator as the PD operation. Then we add a pre-processing on the reconstructed noise map as $\hat{\mathbf{N}}_r = V\mathrm{PD}(\hat{\mathbf{n}})$. The patch-wise noise correlation matrix is calculated on the pre-processed noise patch matrix $\hat{\mathbf{N}}_r$.

V. Experiments

We evaluate the denoising performance of the proposed FINO on both synthetic and real noise with extensive experiments. FINO is implemented using PyTorch, which is tested on a GTX 2080Ti GPU. We adopt the ADAM optimizer with an initial learning rate of 4×10^{-4}. FINO employs two Flow Blocks and twelve coupling layers in each block. Following [31], [38], we utilize a densely connected convolutional block, referred to as Dense Block in [56]. The ratio of clean image and noise variables is fixed as $3:1$, since content contains richer texture and edge information. We set the hyper-parameters $\alpha = 1$, $\beta = 1$, and $\gamma = 0.1$. The network parameters are initialized randomly. During training, we randomly crop patches of resolution 144×144 from input images. We employ Peak Signal-to-Noise (PSNR) for the quantitative evaluation of denoised results. Since the real noisy degradation is relatively mild, causing small PSNR differences in some cases, we also report the Structural SIMilarity (SSIM) for real noise removal experiments.

TABLE I: Comparison of recent denoising methods, including the denoising performance (PSNR) over synthetic noise and real noise, respectively, and computational cost. * denotes the corresponding model is trained with **extra large-scale datasets**, while others are trained with the training set of BSD300 [51].

Method	Synthetic Noise				Real Noise		Computational Cost		
	$\sigma = 20$	$\sigma = 25$	$\sigma = 30$	Avg.	SIDD	DND	Params	Infer time	GFlops
CBM3D [12]	31.27	30.71	28.69	30.22	25.65	34.51	-	-	-
KSVD	-	-	-	-	26.88	36.49	-	-	-
DnCNN [8]	31.68	31.23	28.50	30.47	23.66	32.43	0.56M	-	-
FFDNet	31.48	31.21	29.09	30.59	-	34.40	0.48M	-	-
MIRNet [25]	31.06	30.91	28.33	30.10	-	**39.88**	31.8M	0.8s	196.8
MPRNet [52]	31.32	31.20	28.45	30.32	-	39.80	20.1M	1.5s	548.1
InvDN [31]	30.54	29.56	27.68	29.26	39.28	39.57	2.6M	0.05s	4.8
SwinIR [24]	31.55	31.18	27.81	30.18	-	-	11.5M	1.1s	442.3
SwinIR [24]*	**31.96**	**31.78**	27.76	30.50	-	-	11.5M	1.1s	442.3
FINO	31.82	31.43	**29.56**	**30.94**	**39.40**	39.69	**3.9M**	**0.06s**	**5.5**

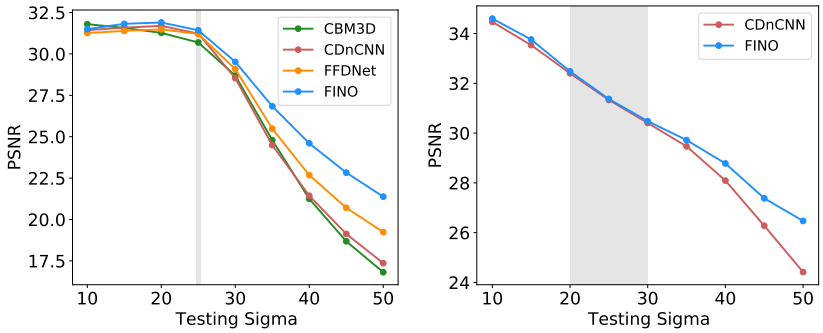


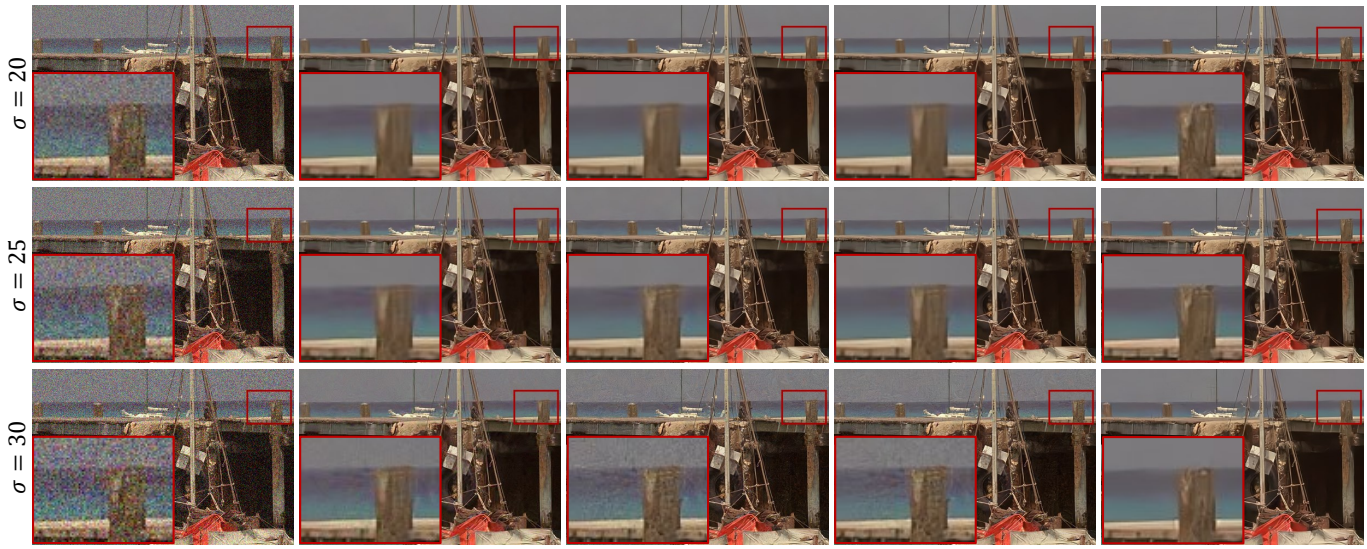
Fig. 4: The generalization capability of the image denoising methods. All the methods are trained over a fixed ranges of noise levels which are indicated by the gray intervals.

A. Synthetic Noise Removal

We simulate spatially invariant additive white gaussian noise (AWGN) with different σ to evaluate the synthetic noise removal performance. We also evaluate if the denoising methods can be generalized to σ that is different from the training corpus. Besides the uniform noise, we further simulate spatially variant AWGN to evaluate the method's robustness.

Spatially invariant AWGN. We evaluate the proposed method in AWGN removal on two widely-used image denoising datasets: CBSD68 [51] and Kodak24 [55]. The CBSD68 dataset consists of 68 images from the separate testing set of the BSD300 dataset [51]. The Kodak24 dataset consists of 24 center-cropped images of size 500×500 from the original Kodak dataset. We only use 200 images selected from the training set of the BSD dataset as training data. The noisy images are obtained by simulating AWGN of noise level $\sigma = 15, 25, 35, 50$ to the clean counterpart. We compare the proposed FINO method with several state-of-the-art denoising methods, including one widely-used classic method (*i.e.,* CBM3D [12]), and deep learning based methods (*i.e.,* CDnCNN [8] and FFDNet [9]). We first test FINO on noisy images corrupted by spatially invariant AWGN. In practice, it is difficult to estimate the noise level accurately, and the estimated noise level can vary in a range. Most existing methods are sensitive to the estimated noise level [57], which means the performance would severely degrade while applied to wrongly estimated noise level. Hence, besides testing the performance when the estimated noise level is accurate, we also test the cases when the noise level is wrongly estimated. For model trained with each estimated σ, we test their performance on $\{\sigma-5, \sigma, \sigma+5\}$ truly sigma variance. The competing methods also are evaluated following the same settings. From the quantitative results shown in Table I and Table III, our method outperforms all competing methods, especially for the wrong σ estimation cases. Although some competing methods are good at removing Gaussian noise when accurate noise estimation, their performance would be significantly degraded once the noise level is wrongly estimated, even slight variation as shown in Figure 4. The examples in Figure 5 also verify the observations in Table III. With the merits of disentanglement and noise model, FINO has strong generalization capability comparing to other methods.

Spatially variant AWGN. We further evaluate the generalization capability of the proposed FINO to deal with spatially variant AWGN. We follow the spatially variant AWGN synthetic approach in [9], which generates a noise level map and applies it to images using element-wise multiplication. We select classic methods (CBM3D [12]) and deep learning based methods (CDnCNN [8], FFDNet [9], and CDnCNN-B (blind version of CDnCNN) as the competing methods.



(a) Noisy (b) CBM3D (c) DnCNN (d) FFDNet (e) FINO (Ours)

Fig. 5: Examples of the denoised results on Kodak24 [55] dataset. From left to right, the input noisy image, the estimated results of CBM3D [12], CDnCNN [8], FFDNet [9], and our method. All the methods are set/trained with the estimated noise level $\sigma = 25$. From top to down, the results for noisy images with $\sigma = 15$, $\sigma = 25$, and $\sigma = 35$.

(a) Noisy (b) CBM3D (c) CDnCNN (d) FFDNet (e) FINO (Ours)

Fig. 6: Examples of the denoised results on CBSD68 [51] dataset. From left to right, the input noisy image, the estimated results of CBM3D [12], CDnCNN [8], FFDNet [9], and our method. All the methods are set/trained with the estimated noise level $\sigma = 50$. From top to down, the results for noisy images with $\sigma = 45$, $\sigma = 50$, and $\sigma = 55$.

In this experiment, we evaluate two versions of FINO: (1) FINO is trained on noisy images with specific $\sigma = 25$; (2) FINO-B is trained on noisy images with a σ range of $(0, 55]$. In the denoising stage, the ground truth noise level map is unavailable. All the methods are set/trained with the estimated noise level $\sigma = 25$, except to the CDnCNN-B, which is trained over a range of noise level $(0, 55]$ and does not need an estimated noise level. The quantitative results are shown in Table II. Our method outperforms all competing methods by large margins, and our blind denoising version also achieves better performances compared with CDnCNN-B. The examples shown in Figure 7 demonstrates our method can coarsely estimate the spatial noise level and effectively reconstruct the clean components, while the other competing methods deliver more over-smoothness or noisy residuum.

B. Real RGB Noise Removal

Finally, we evaluate the performance of different methods on two real-world datasets, *i.e.,* SIDD and DND datasets, which follows a more complicated noise distribution. Real noise stems from multiple sources, *e.g.,* short noise, thermal noise, and dark current noise, and is further affected by the

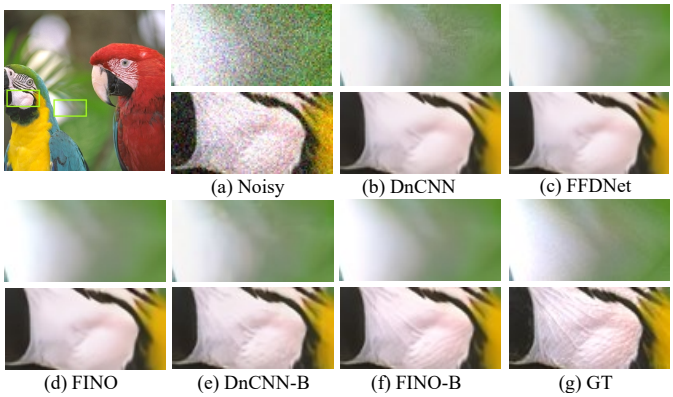


Fig. 7: Generalization capacity of different methods on spatially variant noise removal ($\sigma \in (15, 35)$). All the methods are set/trained with the estimated noise level $\sigma = 25$, except to the blind denoising model, *i.e.,* CDnCNN-B and FINO-B.

TABLE II: Quantitative results on the spatially variant noisy images on Kodak24 [55] dataset. All the methods are set/trained with the estimated uniform noise level $\sigma = 25$, while applying to spatially variant noisy images in the testing stage. Except to the blind denoising methods CDnCNN-B and FINO-B, since they need not estimated noise level.

Method	$\sigma \in (0, 50)$	$\sigma \in (15, 35)$
Noisy	20.66	20.58
CBM3D [12]	29.66	31.16
CDnCNN [8]	29.52	31.41
FFDNet [8]	30.00	31.42
SwinIR [24]	28.78	30.77
SwinIR [24]*	29.06	30.96
FINO (Ours)	**31.15**	**32.01**
CDnCNN-B [8]	31.88	31.75
FINO-B (Ours)	**32.43**	**32.36**

in-camera processing (ISP) pipeline, which can be much more different from uncorrelated noise. We conduct real noise removal on the two most widely-used SIDD [58] and DND [59] datasets. DND does not provide the training set, thus we utilize the medium SIDD dataset as the training set for both evaluating on SIDD and DND, which contains 320 clean and noise image pairs. The performance comparisons on the test set of the SIDD and DND datasets are listed in Table I. The proposed FINO outperforms all competing methods. With the merits of the great generalization capability, the FINO can perform blind real image denoising without an external noise estimation module. Besides, the number of parameters of the FINO is (3.9M), which is much smaller than some competing methods, such as MIRNet (31.8M) and MPRNet (20.1M).

C. Ablation Study

Furthermore, we thoroughly investigate the impact of each loss function applied in the training stage. Table IV shows the evaluation results and Figure 8 demonstrates the visual examples on different combinations of loss functions. Visual

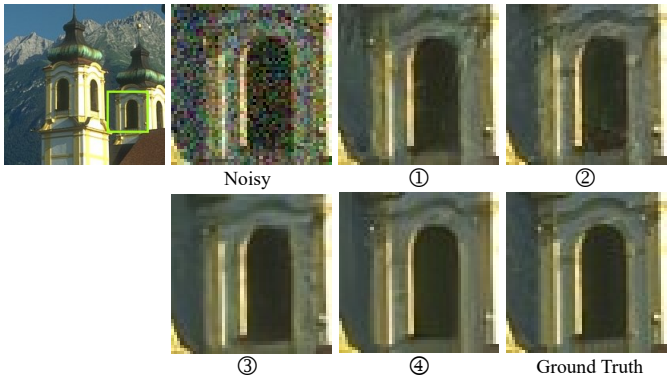


Fig. 8: Visual examples of ablation study, including noisy image, results of four ablation experiments corresponding to the No. in Table IV, and ground truth.

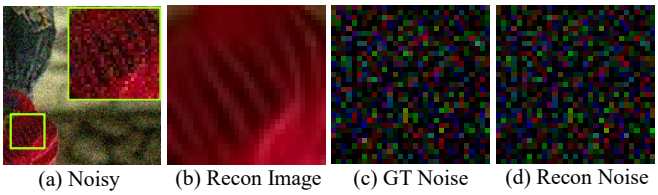


Fig. 9: Visualization of one example from BSD300 dataset with $\sigma = 25$ for reconstructed image and noise. We amplified the real noise (c) and reconstructed noise (d) with $2\times$ for better visibility.

quality of results by FINO without content alignment loss and reconstruction loss drop significantly as shown in the ① and ② in Figure 8. Without the content alignment loss, the generalization capability decreases largely, especially for the robustness for the higher noise as shown in the ① in Table IV. Without the reconstruction loss, the generalization capability would decrease since lacking a strong disentangling constraint. The visual example in Figure 8 also verifies the result with noise loss preserves more structural details.

D. Network Analysis

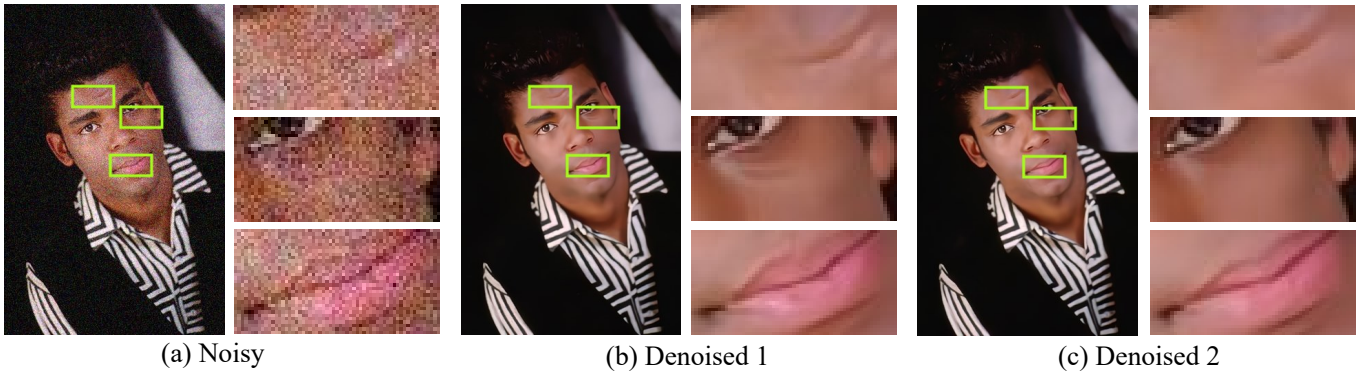
Analysis of one-to-many image denoising. As we mentioned in Section IV-C, the denoised images can also be generated via its disentangled clean image component $\mathbf{z}_c^y$ and a random variable $\mathbf{z}_r$ in the testing stage, as follow

$$\tilde{\mathbf{y}} = \text{Flow}^{-1}(\mathbf{z}_c^y, \mathbf{z}_r) \,. \tag{15}$$

Since the image variable $\mathbf{z}_c$ already contains enough structural information to reconstruct the clean image, the random variable $\mathbf{z}_r$ encoded the bias in the context of image denoising. Thus, $\mathbf{z}_r$ can be randomly sampled a known distribution, *e.g.,* normal distribution, or **zero variable**. Figure 10 demonstrates the denoised results with different sampled $\mathbf{z}_r$. Image denoising is a typical ill-posed inverse problem, which is impossible to reconstruct the exactly clean image with complete information. According to the noisy image, some details are hard to make out, *e.g.,* the length of wrinkle and the shape of the

TABLE III: Quantitative results of denoised PSNR (in dB)↑ on CBSD68 [51], and Kodak24 [55]. The baseline methods include CBM3D [12], CDnCNN [8], and FFDNet [9]. In each column, the best result is highlighted in **bold**.

Datasets	estimated σ	$\sigma = 15$			$\sigma = 25$			$\sigma = 35$			$\sigma = 50$		
	testing σ	10	15	20	20	25	30	30	35	40	45	50	55
CBSD68	CBM3D	34.68	33.52	29.51	31.27	30.71	28.69	29.33	28.89	27.54	27.53	27.38	26.93
	CDnCNN	34.68	33.89	29.06	31.68	31.23	28.50	29.81	29.58	27.63	28.05	27.92	26.83
	FFDNet	34.60	33.87	30.05	31.48	31.21	29.09	29.70	29.58	28.27	28.00	27.96	27.22
	FINO	**34.96**	**34.05**	**30.81**	**31.82**	**31.43**	**29.52**	**29.96**	**29.76**	**28.45**	**28.24**	**28.21**	**27.34**
Kodak24	CBM3D	35.33	34.28	30.21	32.32	31.68	29.68	30.53	29.90	28.52	28.81	28.46	28.12
	CDnCNN	35.23	34.48	29.21	32.28	32.03	28.93	30.57	30.46	28.26	28.98	28.85	27.76
	FFDNet	35.10	34.63	30.47	32.26	32.13	29.81	30.59	30.57	29.15	28.98	28.98	28.22
	FINO	**35.41**	**34.67**	**30.68**	**32.57**	**32.31**	**30.10**	**29.21**	**30.62**	**29.31**	**29.35**	**29.14**	**28.45**



(a) Noisy (b) Denoised 1 (c) Denoised 2

(d) Denoised 3 (e) Denoised 4 (f) GT

Fig. 10: Visual examples of noisy images (a), denoised results (b-e) by FINO with different random variables $\mathbf{z}_r$, and corresponding ground truth (f).

TABLE IV: Quantitative results of ablation study. Evaluated model is trained on $\sigma = 25$ and applied to $\sigma = 20, 25, 30$ for generalization capacity measurements. Note that all variant models include the regression loss $\mathcal{L}_{reg}$.

	$\mathcal{L}_{cnt}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{noise}$	$\sigma = 25$ 20	25	30
①		✓		31.67	31.20	29.17
②	✓			31.66	31.24	29.34
③	✓	✓		31.80	31.32	29.47
④	✓	✓	✓	**31.82**	**31.43**	**29.52**

mouth. Thus, the denoised results may have many different predictions on those areas.

Visualization of noise disentanglement. To better understand the noise model in FINO, we visualize the reconstructed noise and image outputs in Figure 9(b) and (c), respectively. We observe that the reconstructed noise is highly consistent with the ground truth one, demonstrating the effectiveness of FINO's noise modeling.

VI. Limitation and Discussion

Although our FINO performs well in various robust scenarios, it does have limitations in representing complex information, especially in situations where the noise distributions during training and testing are the same. Flow-based networks, unlike traditional CNNs or transformer-based networks, require more rigid architectures. For example, neural flow networks are built by stacking invertible modules like affine coupling layers. This design choice effectively prevents overfitting to training data but also imposes constraints on the

network's ability to represent complex patterns. To achieve further performance gains, one potential option is to utilize a more powerful or larger backbone. However, it is essential to acknowledge that pursuing this enhancement may come with a trade-off, potentially compromising the model's robustness due to an increased risk of overfitting.

VII. Conclusion

In this work, we propose a Flow-based joint Image and NOise model (FINO) to tackle image denoising problems, which aim to estimate the underlying clean image from its noisy measurements. FINO distinctly decouples the image and noise components in the latent space and re-couples them via invertible transformations. Based on that, we employ joint image and noise modeling, *i.e.,* image priors can be learned from the noise-free training corpus, and the noise components are modeled based on the uncorrelated noise assumption. Our experimental results show promising results on both synthetic noise and real noise removal. Furthermore, we demonstrate that our method has superior generalization capability to the removal of non-uniform noise and noise with inaccurate estimation.

References

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2012.

[3] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2862–2869.

[4] H. Liu, R. Xiong, D. Liu, S. Ma, F. Wu, and W. Gao, "Image denoising via low rank regularization exploiting intra and inter patch correlation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 12, pp. 3321–3332, 2017.

[5] H. Wang, Y. Li, Y. Cen, and Z. He, "Multi-matrices low-rank decomposition with structural smoothness for image denoising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 349–361, 2019.

[6] J. Dai, O. C. Au, L. Fang, C. Pang, F. Zou, and J. Li, "Multichannel nonlocal means fusion for color image denoising," *IEEE Transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1873–1886, 2013.

[7] H. Liu, R. Xiong, X. Zhang, Y. Zhang, S. Ma, and W. Gao, "Nonlocal gradient sparsity regularization for image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1909–1921, 2016.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[9] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[10] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 1673–1682.

[11] B. Jiang, Y. Lu, J. Wang, G. Lu, and D. Zhang, "Deep image denoising with adaptive priors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5124–5136, 2022.

[12] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space," in *IEEE International Conference on Image Processing (ICIP)*, vol. 1. IEEE, 2007, pp. I–313.

[13] J. Xu, L. Zhang, W. Zuo, D. Zhang, and X. Feng, "Patch group based nonlocal self-similarity prior learning for image denoising," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 244–252.

[14] L. Guo, Z. Zha, S. Ravishankar, and B. Wen, "Exploiting non-local priors via self-convolution for highly-efficient image restoration," *IEEE Transactions on Image Processing*, vol. 31, pp. 1311–1324, 2022.

[15] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising with block-matching and 3d filtering," in *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, vol. 6064. International Society for Optics and Photonics, 2006, p. 606414.

[16] H. Liu, J. Zhang, and R. Xiong, "Cas: Correlation adaptive sparse modeling for image denoising," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 638–647, 2021.

[17] D. Meng and F. De La Torre, "Robust matrix factorization with unknown noise," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1337–1344.

[18] X. Cao, Y. Chen, Q. Zhao, D. Meng, Y. Wang, D. Wang, and Z. Xu, "Low-rank matrix factorization under general mixture noise distributions," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1493–1501.

[19] H. Zhang, Y. Zhang, L. Zhu, and W. Lin, "Deep learning-based perceptual video quality enhancement for 3d synthesized view," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5080–5094, 2022.

[20] T. Xu, X. Kong, Q. Shen, Y. Chen, and Y. Zhou, "Deep and low-rank quaternion priors for color image processing," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[21] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?" in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2392–2399.

[22] T. Plötz and S. Roth, "Neural nearest neighbors networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[24] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 1833–1844.

[25] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[26] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 683–17 693.

[27] Z. Yue, H. Yong, Q. Zhao, L. Zhang, and D. Meng, "Variational denoising network: Toward blind noise modeling and removal," *arXiv preprint arXiv:1908.11314*, 2019.

[28] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3155–3164.

[29] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1712–1722.

[30] Y. Pan, C. Ren, X. Wu, J. Huang, and X. He, "Real image denoising via guided residual estimation and noise correction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1994–2000, 2022.

[31] Y. Liu, Z. Qin, S. Anwar, P. Ji, D. Kim, S. Caldwell, and T. Gedeon, "Invertible denoising network: A light solution for real noise removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 365–13 374.

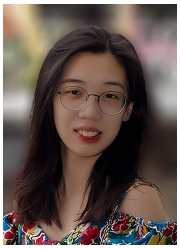
[32] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[33] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

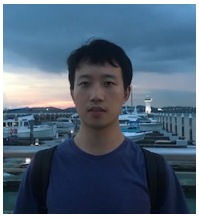
[34] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving flow-based generative models with variational dequantization and architecture design," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2722–2730.

[35] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf

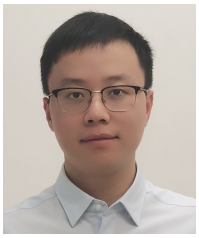
[36] A. Abdelhamed, M. A. Brubaker, and M. S. Brown, "Noise flow: Noise modeling with conditional normalizing flows," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3165–3173.
[37] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 715–732.
[38] M. Xiao, S. Zheng, C. Liu, Y. Wang, D. He, G. Ke, J. Bian, Z. Lin, and T.-Y. Liu, "Invertible image rescaling," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 126–144.
[39] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.
[40] X. Li, X. Jin, J. Lin, S. Liu, Y. Wu, T. Yu, W. Zhou, and Z. Chen, "Learning disentangled feature representation for hybrid-distorted image restoration," in *European Conference on Computer Vision*. Springer, 2020, pp. 313–329.
[41] M. Prabhudesai, S. Lal, D. Patil, H.-Y. Tung, A. W. Harley, and K. Fragkiadaki, "Disentangling 3d prototypical networks for few-shot concept learning," *arXiv preprint arXiv:2011.03367*, 2020.
[42] H. Cheng, Y. Wang, H. Li, A. C. Kot, and B. Wen, "Disentangled feature representation for few-shot image classification," *arXiv preprint arXiv:2109.12548*, 2021.
[43] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 700–708, 2017.
[44] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.
[45] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.
[46] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.
[47] A. H. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, "A unified feature disentangler for multi-domain image translation and manipulation," *arXiv preprint arXiv:1809.01361*, 2018.
[48] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, "Guided image generation with conditional invertible neural networks," *arXiv preprint arXiv:1907.02392*, 2019.
[49] N. Kingsbury and J. Magarey, "Wavelet transforms in image processing," in *Signal analysis and prediction*. Springer, 1998, pp. 27–46.
[50] P. Kirichenko, P. Izmailov, and A. Wilson, "Why normalizing flows fail to detect out-of-distribution data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 2020-December, 2020.
[51] S. Roth and M. J. Black, "Fields of experts: A framework for learning image priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2005, pp. 860–867.
[52] A. Mehri, P. B. Ardakani, and A. D. Sappa, "Mprnet: Multi-path residual network for lightweight image super resolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2704–2713.
[53] Y. Zhou, J. Jiao, H. Huang, Y. Wang, J. Wang, H. Shi, and T. Huang, "When awgn-based denoiser meets real noises," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 074–13 081.
[54] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.
[55] R. Franzen, "Kodak lossless true color image suite," *source: http://r0k.us/graphics/kodak*, vol. 4, no. 2, 1999.
[56] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.
[57] S. Mohan, Z. Kadkhodaie, E. P. Simoncelli, and C. Fernandez-Granda, "Robust and interpretable blind image denoising via bias-free convolutional neural networks," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJlSmC4FPS
[58] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
[59] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1586–1595.



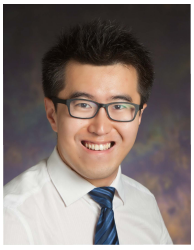
Lanqing Guo received the B.Eng. degree in software engineering from Wuhan University, China. She is currently a Ph.D. candidate in the School of Electrical and Electronic Engineering at Nanyang Technological University (NTU), Singapore. Her current research interests focus on image processing, computational imaging, and computer vision.



Siyu Huang received the B.E. degree and Ph.D. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2014 and 2019. He is currently an Assistant Professor with the School of Computing, Clemson University. Before that, he was a Visiting Scholar at Language Technologies Institute in the School of Computer Science, Carnegie Mellon University in 2018, a Research Scientist at Big Data Laboratory, Baidu Research from 2019 to 2021, a Research Fellow in the School of Electrical and Electronic Engineering at Nanyang Technological University in 2021, and a Postdoctoral Fellow in the John A. Paulson School of Engineering and Applied Sciences at Harvard University from 2021 to 2023. He has published 30 papers on top-tier computer science journals and conferences. His research interests are primarily in computer vision, deep learning, and generative AI.



Haosen Liu received the B.S degree and the M.S. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, the University of Hong Kong, Pokfulam, Hong Kong. His research interests include image processing and computer vision.



Bihan Wen received the received the B.Eng. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2015 and 2018, respectively. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include machine learning, computational imaging, computer vision, image and video processing, and big data applications. Dr. Wen was a recipient of the 2016 Yee Fellowship and the 2012 Professional Engineers Board Gold Medal. He was also a recipient of the Best Paper Runner Up Award at the IEEE International Conference on Multimedia and Expo in 2020. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He has served as a Guest Editor for IEEE SIGNAL PROCESSING MAGAZINE in 2022, and IEEE Journal of Selected Topics in Signal Processing in 2023. He is an IEEE Senior Member.