Deep Variational Network Toward Blind Image Restoration

Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, Fellow, IEEE, Deyu Meng, Member, IEEE, and Kwan-Yee K. Wong, Senior, IEEE

Abstract—Blind image restoration (IR) is a common yet challenging problem in computer vision. Classical model-based methods and recent deep learning (DL)-based methods represent two different methodologies for this problem, each with its own merits and drawbacks. In this paper, we propose a novel blind image restoration method, aiming to integrate both the advantages of them. Specifically, we construct a general Bayesian generative model for the blind IR, which explicitly depicts the degradation process. In this proposed model, a pixel-wise non-i.i.d. Gaussian distribution is employed to fit the image noise. It is with more flexibility than the simple i.i.d. Gaussian or Laplacian distributions as adopted in most of conventional methods, so as to handle more complicated noise types contained in the image degradation. To solve the model, we design a variational inference algorithm where all the expected posterior distributions are parameterized as deep neural networks to increase their model capability. Notably, such an inference algorithm induces a unified framework to jointly deal with the tasks of degradation estimation and image restoration. Further, the degradation information estimated in the former task is utilized to guide the latter IR process. Experiments on two typical blind IR tasks, namely image denoising and super-resolution, demonstrate that the proposed method achieves superior performance over current state-of-the-arts. The source code is available at https://github.com/zsyOAoA/VIRNet.

Index Terms—Image restoration, denoising, super-resolution, generative model, variational inference.

1 INTRODUCTION

Image restoration (IR) is an active research topic in the fields of signal processing and computer vision. It aims at recovering the latent high-quality image \( z \) from the observed corrupted counterpart \( y \), i.e.,

\[
y = Hz + n,
\]

where \( H \) is the degradation operator, and \( n \) is image noise. With different degradation settings for \( H \), Eq. (1) represents different IR tasks. For example, the classical IR tasks, such as image denoising, deblurring, and super-resolution, can be easily obtained by setting \( H \) as an identity matrix, a blurring operator, and a composition of blurring and downsampling operators, respectively. The difficulties of IR tasks mainly come from \( H \) and \( n \). The former inclines to cause severe information loss in some tasks, like deblurring and super-resolution, and the latter is usually complicated due to the accumulation of noises from multiple sources e.g., capturing instrument, camera pipeline and image transmission [1]. In blind IR tasks, we need to simultaneously solve the problems of degradation estimation and image restoration, which makes it more challenging.

In the past decades, plenty of IR methods have been proposed under the maximum a posteriori (MAP) framework. From the Bayesian perspective, it generally involves a likelihood term and a prior term. More specifically, the likelihood term encodes the image degradation process of Eq. (1), while the prior term reflects our subjective knowledge of the latent high-quality image. Most of these methods mainly focused on designing more effective image priors so as to alleviate the ill-posedness of IR tasks. Commonly used image priors include total variation (TV) [2], non-local similarity [3], [4], sparsity [5]–[7], low-rankness [8]–[10] and so on. In contrast, other works focused on the likelihood term by constructing more flexible noise distributions, e.g., mixture of Gaussian (MoG) [11], mixture of Exponential (MoEP) [12], and Dirichlet Process mixture of Gaussian (DP-MoG) [13]–[15]. Even though these model-based methods are with highly intuitive physical meanings and also generalize well in most scenarios, they still have evident defects. Firstly, these methods are always time-consuming, since they require to re-solve the whole model for any new testing images. Such a one-by-one optimized paradigm tends to bring up large computational burdens, making them very hard to be applied in real applications. Secondly, limited by the manually designed likelihood and image priors, which usually cannot faithfully represent the image knowledge, they struggle to handle some complex modeling problems in real cases, such as the blind IR tasks with complicated image degradation.

Different from the aforementioned model-based methods, current deep learning (DL)-based methods represent another research trend. Their core idea is to employ the deep
neural networks (DNNs), being with powerful fitting capability, to directly learn the image knowledge from a large amount of pre-collected image pairs in an end-to-end training manner. Dong et al. [16] and Zhang et al. [17] firstly proposed SRCNN and DnCNN that surpassed classical model-based methods in image super-resolution and denoising, respectively. Subsequently, many DL-based methods [18]–[26] were proposed and they achieved unprecedented successes in the field of IR. While they have achieved huge boost in performance, most of them ignore the modeling mechanism underlying the image degradation, especially the image noise. To be specific, the $L_2$ or $L_1$ loss commonly used in current DL-based methods indeed implies that the noise $n$ in Eq. (1) obeys the independent and identically distributed (i.i.d.) Gaussian or Laplacian distribution [11], [12]. This, however, always deviates from true noise configurations in real cases. For example, the camera sensor noises in practical image denoising are signal-dependent and spatially variant, and thus evidently non-i.i.d. in statistics. Neglecting such intrinsic noise properties will certainly harm the generalization capability of the model in real scenarios with complicated noises.

As analyzed above, the model-based methods are capable of encoding the image degradation through the likelihood, but hindered by the limited model capacity and the slow inference speed. In contrast, the DL-based methods, equipped with DNNs, are with large model capacity and powerful non-linear fitting capability. What’s more, in the testing phase, these methods are much faster than the model-based method, since they only need one feedforward pass for any newly coming image. This naturally inspires us to develop a new IR method, which is expected to combine both the advantages of the classical model-based methods and recent DL-based methods. In this work, we take one step forward along this research line by proposing a deep variational model for blind IR. It first constructs a traditional probabilistic model for IR, and then embeds the powerful DNNs into its posterior inference to increase the model capacity. Specifically, the contributions of this work can be summarized from two aspects, namely model construction and algorithm design, as follows:

On one hand, a Bayesian generative model is built for general IR tasks, and thus naturally inherits the advantages of classical model-based methods. Furthermore, we also consider a more complicated degradation process when building our model in this study:

- Different from the commonly used independent factorization strategy in VI, we factorize the expected posterior, namely the joint distribution of the degradation information and the latent clean image, into a conditional form. Such a formulation derives a unified framework to simultaneously deal with the tasks of degradation estimation and image restoration, in which the degradation information estimated by the former provides sound guidance for the subsequent restoration task.

- To largely increase the fitting capability of our model, the desirable posteriori distributions are parameterized by DNNs, and then optimized in an amortized manner during training. In the testing phase, the well-trained model is capable of fastly inferring the posteriori distribution of any new testing image in an explicit manner, and thus evidently more efficient than the classical model-based methods.

In summary, this work aims to explore a novel modeling paradigm, which is expected to integrate the merits both of the classical model-based methods and recent DL-based methods, for the IR problem. A preliminary version of this work has been published in NeurIPS 2019 [27] which focuses only on image denoising. This present work makes substantial improvements on model construction, the inference algorithm, and the empirical evaluations over the conference version. Especially, we consider a more general degradation process (i.e., Eq. (1)) to build the Bayesian generative model, making it capable of handling more complicated and general IR tasks, such as blind image super-resolution.

The remainder of the paper is organized as follows: Section 2 introduces the related work. Section 3 proposes our generative model, and discusses two typical IR tasks. Section 4 presents the designed stochastic VI algorithm for solving our model. In Section 5, experiments are demonstrated to evaluate the performance of our method. Section 6 finally concludes the paper.

## 2 Related Works

In this section, we first review model-based and DL-based IR methods. We then briefly summarize recent explorations that attempts to combine both of these two methodologies.

### 2.1 Model-based Methods

Most of the classical model-based methods can be formulated into the MAP framework, which contains a likelihood (fidelity) term and a prior (regularization) term from the Bayesian perspective. Relevant developments thus mainly focused on these two terms.

#### Prior Modeling Methods

Aiming at alleviating the ill-posed issue of IR, many studies attempted to exploit rational image prior knowledge. Statistical regularities exhibited in images were firstly exploited, e.g., TV denoising [2] and wavelet coring [28]. Then, NLM [3] and BM3D [4] were both proposed for denoising based on the non-local self-similarity prior, meaning that small image patches in a non-local image area possess similar configurations. Later, low-rankness [8]–[10] and sparsity [4]–[7] priors, which also aim to explore the characteristics of image patches, became
popular and were widely used in IR tasks. To further increase the model’s capacity and expression ability, some other methods moved from analytical technologies to data-driven approaches. E.g., Roth and Black [29] proposed the fields of experts (FoE) to learn image priors. Barbu [30] trained a discriminative model for the Markov random field (MRF) prior, while Sun and Tappen [31] proposed a non-local range MRF (NLR-MRF) model. More related works can be found in [32], [33].

**Noise Modeling Methods.** Different from the prior modeling methods, noise modeling methods focus on the likelihood (fidelity) term of the MAP framework. In fact, the widely used $L_1$ or $L_2$ loss functions implicitly make the i.i.d. Gaussian or Laplacian assumptions on image noise, which often underestimates the complexity of real noise. Based on this observation, Meng et al. [11] proposed the MoG noise modeling method under the low-rankness framework. Furthermore, Zhu et al. [13], [34] and Yue et al. [14], [15] both introduced the non-parametric Dirichlet Process into MoG to increase its flexibility, leading to the adaptive adjustment for the component number of MoG.

### 2.2 DL-based Methods

DL-based methods represent a data-driven trend for the IR task. They straightforwardly train an explicit mapping function parameterized by DNN on a large amount of image pairs. The earliest convolution neural network (CNN) method can be traced back to [35], in which a five-layer network was employed. Some auto-encoder-based methods were then proposed in [36], [37]. Due to the insufficient research on DL technologies, however, these methods were inferior to the model-based methods in performance.

The first significant improvement of DL-based methods was achieved by [38] which obtained comparable performance with BM3D [4] in denoising task using a plain multi-layer perceptron. Subsequently, Dong et al. [16] proposed the first CNN model for super-resolution, and it outperformed the classical model-based methods. With the advances of deep learning, Zhang et al. [17] trained a deep CNN model named DnCNN and achieved state-of-the-art performance in denoising, JPEG deblocking, and super-resolution. Since then, the DL-based methods began to dominate the research trend in almost all of the IR tasks, especially in denoising [18], [20], [23], [25], [39] and super-resolution [40]–[46].

Inspired by the development of generative adversarial network (GAN) [47], some DL-based methods also followed the research line of noise modeling introduced in Sec. 2.1. Typically, Chen et al. [48] proposed a noise generator to simulate the real noise under the adversarial training mechanism, and Kim et al. [49] further introduced some camera settings (e.g., ISO level and shutter speed) into the generator as extra guidance. More recently, Yue et al. [24] proposed a dual adversarial loss to implement the noise removal and noise generation tasks into one unique Bayesian framework. Different from these GAN-based implicit noise modeling manners, this study adopts a more powerful and flexible non-i.i.d. Gaussian distribution to explicitly model the image noise, which avoids the instability in training GAN.

### 2.3 Some Relevant Explorations

The research to combine the model-based methods and the DL-based methods has attracted increasing attention in recent years. Some significant explorations have been attempted towards this goal.

Deep plug-and-play methods [50]–[54] usually replace the denoising sub-problem in model-based methods with one or multiple pre-trained DNN denoisers, to leverage the abundant image prior knowledge learned by such deep denoisers. Due to the lack of end-to-end training, they always rely on tedious hyper-parameters tuning to guarantee stable performance. To alleviate this drawback, deep unfolding methods [55]–[59] take a step forward by embedding the DNNs into traditional optimization algorithms (e.g., HQS, ADMM). Attributed to the end-to-end training manner, these deep unfolding methods achieve promising performance in some IR tasks.

Deep image prior (DIP) [60] and its related methods [61]–[64] represent another significant approach along this research line. These methods aim to seek a corresponding DNN model that maps the pre-sampled noise to the desirable clean image for any given corrupted image under the MAP framework. Similar to the model-based methods, they are mainly limited by the time-consuming optimization process in the testing phase.

In this study, we develop a general and novel deep IR model which is evidently different from the above research approaches. From the model perspective, a Bayesian generative model is constructed for general IR tasks, naturally inheriting the capability of modeling image degradation from the classical model-based methods. From the algorithm perspective, we design an amortized VI algorithm that parameterizes all the involved posterior distributions with DNNs. The embedded DNNs in our algorithm equip it with powerful fitting ability and fast testing speed as recent DL-based methods.

### 3 The Proposed Method

#### 3.1 Basic Settings

In this paper, we consider two commonly used settings on the degradation operator $H$ of Eq. (1). In the first case, $H$ is an identity matrix, corresponding to the task of image denoising. The difficulties of this task are naturally attributed to the complexity of the image noise, which is often spatially variant and signal-dependent in real scenarios. Besides, their statistics (e.g., the noise levels) are always unknown in blind image denoising. It is thus necessary to devise methods to estimate the noise distribution and recover the latent high-quality image simultaneously.

In the second case, we consider a more general IR task, namely image super-resolution, in which $H$ is a composition of blurring and downsampling. The downsampling operation leads to serious information loss, especially in the case of large scale factors, making it more challenging compared with denoising. Similarly, blind super-resolution also involves two sub-tasks, i.e., estimating the degradation information, including both the blur kernel and noise distribution, and restoring the high-resolution image.

In addition, we briefly introduce some necessary settings on the training and testing data. The training data consists of
multiple triplets, i.e., \( D = \{ y^{(j)}, x^{(j)}, H^{(j)} \}_{j=1}^N \), where \( y^{(j)} \) and \( x^{(j)} \) denote the corrupted image and the underlying high-quality image, respectively. \( H^{(j)} \) represents the degradation operator, which is an identity matrix for denoising and a blur kernel for super-resolution. The superscript \( j \) of \( H^{(j)} \) indicates that the degradation model varies from one sample to another in our training data. It should be noted that, in the real-world image denoising task, \( x^{(j)} \) is usually estimated by averaging several noisy images taken under the same camera conditions [65]. In the testing phase, given only one corrupted image, our goal is to first estimate the degradation information, and then recover the high-quality image based on such information.

Next, we formulate a rational Bayesian generative model for the tasks of denoising and super-resolution.

### 3.2 Model Formulation for Denoising

Let \( \{ y, x \} \in D \) denote any noisy/noise-free image pair in the training dataset. For the noisy image \( y \), we assume that it is generated as follows:

\[
y_i \sim \mathcal{N}(y_i | z_i, \sigma^2), \quad i = 1, 2, \ldots, d,
\]

where \( \mathcal{N}(\cdot | \mu, \sigma^2) \) denotes the Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \). \( z \) denotes the latent clean image, \( d \) is the number of pixels in the noisy image. Notably, different from the commonly used i.i.d. Gaussian/Laplacian assumption, we model the image noise as pixel-wise non-i.i.d. Gaussian distribution in Eq. (2). Such a non-i.i.d. noise assumption largely increases the degrees of freedom of the noise distribution and is thus expected to better fit complicated real noise as illustrated in Sec. 5.3.1.

Next, we introduce some prior knowledge for the latent clean image \( z \) and the noise variance map \( \sigma^2 \). In the real dataset, \( x \) provides an approximate estimation to the underlying clean image \( z \) by some statistical method [65]. Therefore, it can be embedded into the following prior distribution as a constraint for \( z \):

\[
z_i \sim \mathcal{N}(z_i | x_i, \sigma^2_0), \quad i = 1, 2, \ldots, d,
\]

where \( \varepsilon_0 \) is a hyper-parameter that reflects the closeness between \( x \) and \( z \). In some synthetic experiments where the underlying clean image is accessible, \( x \) is indeed the true clean image \( z \), which can be easily formulated by setting \( \varepsilon_0 \) as a small number close to 0. Under this setting, Eq. (5) degenerates to the Dirac distribution centered at \( x \).

As for the variance map \( \sigma^2 \), we construct the following inverse Gamma distribution as its conjugate prior [66]:

\[
\sigma^2_i \sim \text{IG} \left( \sigma^2_i | \alpha_0 - 1, \alpha_0 \xi_i \right), \quad i = 1, 2, \ldots, d,
\]

where

\[

\xi_i = \mathcal{G} \left( (\hat{y} - H \hat{x})^2 ; \rho \right).
\]

\( \mathcal{G}(\cdot ; \rho) \) is the Gaussian filter with window size \( p \), \( \hat{y}, \hat{x} \in \mathbb{R}^{h \times w} \) are the matrix (image) forms of \( y, x \in \mathbb{R}^d \), and \( \alpha_0 \) is the shape parameter of the inverse Gamma distribution.

Actually, \( \xi_i \) in Eq. (5) provides an estimation for the variance \( \sigma^2_i \). It is calculated using a Gaussian filter in the \( p \times p \) window centered at the \( i \)-th pixel. The elaborate design on the shape and scale parameters (i.e., \( \alpha_0 - 1 \) and \( \alpha_0 \xi_i \)) in Eq. (4) guarantees that the mode of this prior distribution is \( \xi_i \) exactly. The hyper-parameter \( \alpha_0 \) controls the strength of this prior distribution, and it is set as \( p^2/2 \) in our method following [67]. More explanations on this hyper-parameter can be found in Appendix A.

### 3.3 Extension to Blind Super-resolution

For the problem of image super-resolution, the degradation model in Eq. (1) can be reformulated as:

\[
y = (z * k) \downarrow_s + n,
\]

where \( k \) denotes the blur kernel, \( * \) is the convolution operator, and \( \downarrow_s \) is the downsampler with scale factor \( s \). Based on this degradation model, we extend the noise assumption of Eq. (2) as follows:

\[
y_i \sim \mathcal{N}(y_i | (z * k) \downarrow_s i, \sigma^2_i), \quad i = 1, 2, \ldots, d,
\]

where \( [x]_i \) represents the \( i \)-th pixel of \( x \), i.e., \( x_i \).

To handle blind super-resolution, the most challenging part is how to model the blur kernel. Recently, lots of related literature [57, 63, 68–70] have found that the anisotropic Gaussian kernels are sufficient to guarantee pleasing results for image super-resolution. In this study, we follow these related works and adopt the anisotropic Gaussian kernels.

Thus, the blur kernel \( k \) with size \( (2r + 1) \times (2r + 1) \) can be defined as:

\[
k_{ij} = g(\lambda_1^2, \lambda_2^2, \rho) = \frac{1}{2\pi\lambda_1\lambda_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2} S^T \Sigma^{-1} S \right\},
\]

where \( \Sigma = \begin{bmatrix} \lambda_1^2 & \lambda_1\lambda_2\rho \\ \lambda_1\lambda_2\rho & \lambda_2^2 \end{bmatrix} \) is the covariance matrix, \( \rho \) is the Pearson correlation coefficient, \( S = [i, j]^{T} \) denotes the spatial coordinate with \( i, j \in \{-r, \ldots, r\} \). By denoting \( A = \{\rho, \lambda_1^2, \lambda_2^2\} \), one can easily observe that the blur kernel \( k \) is completely determined by \( A \) when the kernel size is fixed. This inspires us to design a prior distribution for \( A \) instead of \( k \).

In essence, Eq. (8) represents the blur kernel \( k \) through two Gaussian distributions with variance parameters \( \lambda_1^2 \) and \( \lambda_2^2 \) along the horizontal and vertical directions, and their correlations is depicted by \( \rho \). For \( \{\lambda_1, \lambda_2\} \) and \( \rho \), we impose
the inverse Gamma [66] and Dirac distributions for them as prior constraints, respectively, i.e.,

\[
A \sim p(A) = \text{Dirac}(\rho | \hat{\rho}) \prod_{l=1}^{2} IG(\lambda_1^2 | \kappa_0 - 1, \kappa_0 * \hat{\lambda}_1^2),
\]

(9)

where \(\{\rho, \hat{\lambda}_1^2, \hat{\lambda}_2^2\}\) reflect the corresponding true kernel information contained in the training data. Similar to the \(\alpha_0\) of Eq. (4), \(\kappa_0\) is also a hyper-parameter controlling the shape of the inverse Gamma distributions. For the purpose of easy optimization, we employ a Gaussian distribution with a small variance to approximate the Dirac distribution. Thus, based on Eq. (8), the blur kernel \(k\) is modelled as:

\[
k = q(A),
\]

(10)

\[
A \sim p(A) = \mathcal{N}(\rho | r_0^2 \hat{\rho}, r_0^2 \beta),
\]

(11)

where the variance \(r_0^2\) is empirically set as 1e-4 throughout all our experiments.

Combining Eqs. (3)-(5), (7), and (10)-(11) (or Eqs. (2)-(5)) together, a full Bayesian model for blind image super-resolution (or image denoising) can be obtained. The corresponding graphical model is depicted in Fig. 1. Our goal then turns to infer the posterior of latent variables \(\{z, \sigma^2, A\}\) (or \(\{z, \sigma^2\}\)) from \(y\), namely \(p(z, \sigma^2, A | y)\) (or \(p(z, \sigma^2 | y)\)).

4 Stochastic Variational Inference

In this section, a stochastic variational inference algorithm is designed for the proposed generative model. In the following part, we take blind super-resolution problem as an example to present our algorithm, since it can be easily degenerated into the denoising task by setting the blur kernel as Dirac delta function and the scale factor as 1.

4.1 Form of Variational Posterior

Inspired by the VI techniques [66], we firstly construct a variational distribution \(q(z, \sigma^2, A | y)\) to approximate the true posterior \(p(z, \sigma^2, A | y)\) led by our generative model. The variational posterior distribution is then conditionally factorized as:

\[
q(z, \sigma^2, A | y) = q(z | y, \sigma^2, A)q(\sigma^2 | y)q(A | y).
\]

(12)

Next, we begin to design specific forms for these three factorized posteriori distributions.

The conjugate prior of Eq. (4) inspires us to assume \(q(\sigma^2 | y)\) as the following inverse Gamma distribution:

\[
q(\sigma^2 | y) = \prod_{i=1}^{d} IG(\alpha_0^2 | \alpha_0 - 1, \alpha_0 \beta(y; W_S)),
\]

(13)

where \(\beta_i(y; W_S)\) is a mapping function parameterized as a DNN named as sigma network (SNet) with parameters \(W_S\). It aims to predict the scale parameter of \(q(\sigma^2 | y)\) directly from the corrupted image \(y\). As for the shape parameter of \(q(\sigma^2 | y)\), we simply fix it as the same to the prior distribution, i.e., \(\alpha_0 - 1\). Different from the strategy of setting them both as learnable parameters in our previous version [27], such a modification largely simplifies the evidence lower bound of Sec. 4.2 and also makes our algorithm more stable during training. Similarly, we formulate \(q(A | y)\) as:

\[
q(A | y) = \mathcal{N}(\rho | m(y; W_K), r_0^2) \prod_{l=1}^{2} IG(\lambda_1^2 | \kappa_0 - 1, \kappa_0 * \hat{\lambda}_1^2),
\]

(14)

where \(m(y; W_K)\) and \(\eta(y; W_K)\) are jointly parameterized as a DNN with parameters \(W_K\), named as kernel network (KNet). It takes the low-resolution image \(y\) as input and outputs the posteriori parameters for \(q(A | y)\).

As for \(q(z | y, \sigma^2, A)\), we set it as Gaussian distribution:

\[
q(z | y, \sigma^2, A) = \prod_{i}^{d} \mathcal{N}(z_i | \mu_i(y, \sigma^2, A; W_R), \epsilon_0^2),
\]

(15)

where \(\mu_i(y, \sigma^2, A; W_R)\) represents the mapping function to evaluate the mean value of the posteriori Gaussian distribution of \(z\). Naturally, it is parameterized as a DNN with parameters \(W_R\), named as restoration network (RNet). For the ease of training, we set the variance parameter of this posteriori distribution as a constant, i.e., \(\epsilon_0^2\), being the same with that of the prior distribution in Eq. (3).
It is necessary to emphasize that the posteriori distribution $q(z|y, \sigma^2, A)$ is conditioned on $\sigma^2$ and $A$, which means that $RNet$ depends on the noise variance map estimated by $SNet$ and the kernel information predicted by $KNet$. Generally speaking, the conditional assumption of Eq. (12) decomposes the task of blind super-resolution into two cascaded sub-tasks, namely degradation estimation implemented by $SNet$ and $KNet$, and non-blind image restoration implemented by $RNet$. The whole inference procedure is shown in Fig. 2.

Remark. In Eq. (13), the mode of $q(\sigma^2|y)$ is just equal to $\beta(y; W_S)$, which is predicted by $SNet$. In other words, we leverage $SNet$ to only estimate the core posterior parameter, namely the mode, instead of the whole posteriori distribution. The reasons underlying this setting are two-fold: on one hand, this strategy inclines to alleviate the learning burden of $SNet$ to some extent. On the other hand, we can directly utilize the output of $SNet$ as an estimated variance map to solve some downstream problems that depend on pre-known noise levels. Similarly, we also employ this partial learning strategy in Eq. (14) and Eq. (15).

4.2 Evidence Lower Bound

In this part, we first induce a rational objective function, namely the evidence lower bound (ELBO), to train our model. Then, we give a further discussion to explain the relationship between our designed ELBO and the traditional mean-square error (MSE) loss.

4.2.1 Derivation of ELBO

For the convenience of presentation, we simply denote $\beta_i(y; W_S), m(y; W_K), \eta_i(y; W_K)$, and $\mu_i(y, \sigma^2, A; W_R)$ as $\beta_i, m, \eta_i$, and $\mu_i$, respectively. Given any corrupted image $y$, its logarithm marginal probability can be decomposed as

$$
\log p(y) = \mathcal{L}(z, \sigma, A; y) + KL[q(z, \sigma^2, A|y)||p(z, \sigma^2, A|y)],
$$

where

$$
\mathcal{L}(z, \sigma, A; y) = E_q[\log p(y|z, \sigma^2, A)p(z)p(\sigma^2)p(A)] - \log q(z, \sigma^2, A; y).
$$

Here $E_q[\cdot]$ denotes the expectation w.r.t. the posteriori distribution $q(z, \sigma^2, A|y)$. The second term of Eq. (16) represents the KL divergence between the variational posterior $q(z, \sigma^2, A|y)$ and the true posterior $p(z, \sigma^2, A|y)$. Due to the non-negativeness of KL divergence, $\mathcal{L}(z, \sigma, A; y)$ constitutes a lower bound of $\log p(y)$, thus called ELBO. Therefore, we can naturally approximate the true posteriori $p(z, \sigma^2, A|y)$ with $q(z, \sigma^2, A|y)$ via maximizing ELBO.

Combining the factorized assumption of Eq. (12), the ELBO can be rewritten as

$$
\mathcal{L}(z, \sigma, A; y) = E_q(\sigma, A|y) \log p(y|z, \sigma^2, A) - E_q(\sigma, A|y) KL[q(z|y, \sigma^2, A)||p(z)] - KL[q(\sigma^2|y)||p(\sigma^2)] - KL[q(A|y)||p(A)],
$$

where $q(\sigma, A|y) = q(\sigma|y)q(A|y)$.

Next, we consider how to calculate each term in Eq. (18) step by step. The first term is intractable, mainly because the posterior $q(z, \sigma, A|y)$ is parameterized as complicated forms of DNNs. Fortunately, we can use the reparameterization trick [71] to obtain multiple differentiable samples from the posteriors, and then use them to estimate these two terms by Monte Carlo (MC) like VAE [71]. Concretely, the re-sampling process from $q(z|y, \sigma^2, A)$ can be easily implemented as

$$
\bar{z} = \mu + \varepsilon_0 \tau, \tau \sim N(\tau|0, I).
$$

In the training phase, we sample from $q(A|y)$ and $q(\sigma^2|y)$ utilizing the pathwise derivative technology [72], and denote the re-sampled data example as $\tilde{A}$ and $\tilde{\sigma}^2$. Based on $\tilde{z}$, $\tilde{A}$ and $\tilde{\sigma}^2$, the first term of Eq. (18) can be approximated as follows:

$$
E_q(z, \sigma, A|y)[\log p(y|z, \sigma^2, A)] \approx -\frac{1}{2} \sum_{i=1}^{d} \left\{ \log \tilde{\sigma}_i + w_i \left( y_i - \left( \tilde{z} \cdot \tilde{k} \right)_{\downarrow i} \right)^2 \right\},
$$

where $w_i = \frac{1}{\bar{\sigma}^2}, \tilde{k} = g(\tilde{A})$, and $g(\cdot)$ is defined in Eq. (10). Note that we have omitted a constant that is independent of the learnable parameters in Eq. (20).

As for the last three terms of Eq. (18), they can all be analytically calculated as follows:

$$
KL[q(z|y, \sigma^2, A)||p(z)] = \sum_{i=1}^{d} (\mu_i - x_i)^2 / 2\hat{\sigma}_i, \quad KL[q(\sigma^2|y)||p(\sigma^2)] = \sum_{i=1}^{d} \alpha_0 \left( \frac{\xi_i}{2\hat{\sigma}_i} + \log \frac{\beta_i}{\hat{\sigma}_i} - 1 \right), \quad KL[q(A|y)||p(A)] = \frac{(m - \tilde{\mu})^2}{2\hat{\sigma}_A} + \sum_{i=1}^{2} \kappa_0 \left( \frac{\tilde{\lambda}_i}{\hat{\eta}_i} + \log \frac{\eta_i}{\hat{\lambda}_i} - 1 \right).
$$

Finally, we can get the expected objective function, namely the negative ELBO on the entire training dataset, to optimize the network parameters of $W_S, W_K$, and $W_R$ as follows:

$$
\min_{W_S, W_K, W_R} \sum_{j=1}^{N} \mathcal{L}(z^{(j)}, \sigma^{(j)}, A^{(j)}; y^{(j)}),
$$

where $z^{(j)}, \sigma^{(j)},$ and $A^{(j)}$ denote the posteriori parameters for the $j$-th image pair in training dataset $D$.

In the testing phase, we first predict the variance parameter $\beta$ as well as the kernel parameters $(m, \eta_1, \eta_2)$ using $SNet$ and $KNet$. Subsequently, these predicted parameters are concatenated with the low-resolution image $y$ and serve as the input of $RNet$. The output of $RNet$, i.e., the mean parameter $\mu$ of the posterior of the underlying high-quality image, can be reasonably regarded as our desirable restored result. The whole inference pipeline is illustrated in Fig. 2.

4.2.2 Discussion

Let us delve further into the ELBO loss of Eq. (18). The first term of this formulation presents the log-likelihood of the observed low-resolution image, enforcing that the recovered high-resolution image can be mapped back to the low-resolution one via the estimated degradation model. The other three terms correspond to the constraints for
the estimated high-resolution image $z$, the noise variance $\sigma^2$, and the kernel parameters $\Lambda$, respectively, as shown in Eqs. (21)-(23). Particularly noteworthy is the constraint imposed on $z$ in Eq. (21), which takes the form of an MSE endowed with a weight factor of $\frac{1}{\sigma^2}$. In other words, our designed ELBO loss degenerates into the commonly used MSE by setting the hyper-parameter $\varepsilon_0$ to a small value close to zero. A more thorough experimental comparison with the MLE baseline can be found in Sec. 5.4.

We now specifically consider the log-likelihood term, commonly called as data fidelity within the field of IR. Most of the existing IR methods [8], [9] usually posit uniform importance to each element in this term, i.e., $\sum_i \left(y_i - \left(\hat{z} \cdot \hat{k}\right) \downarrow\circ I_i\right)^p$. This study introduces an innovative approach whereby we exploit an adaptive manner to re-weight the data fidelity in terms of $l_2$-norm, i.e., $\sum_i w_i \left(y_i - \left(\hat{z} \cdot \hat{k}\right) \downarrow\circ I_i\right)^2$, as detailed in Eq. (20). Each pixel is assigned with a weight of $w_i = \frac{1}{\sigma^2}$, in which $\hat{\sigma}^2$ is sampled from the noise distribution estimated by SNet. This re-weighting strategy, based on noise variance, aligns with the principles of Bayesian statistics, akin to prior literature such as [11], [13].

4.3 Network Structure and Learning

As shown in Fig. 2, SNet takes the corrupted image $y$ as input and outputs the scale-related parameter of $q(\sigma^2|y)$, achieving the goal of noise estimation. In practice, it consists of five convolution layers, and each is followed with a Leaky ReLU activation except for the first and last layers. As for the KNet, it is designed to predict the posteriori distribution of the kernel parameter $\Lambda$ from the corrupted image $y$. In the implementation, we firstly employ one convolutional layer and eight channel attention blocks (CAB) [23] to extract abundant feature maps, and then fuse them by one convolutional layer followed by an average pooling layer to obtain the posteriori parameters in $q(\Lambda|y)$.

The design of RNet, aiming to infer the conditional posteriori distribution of the desirable high-quality image, plays the most important role in blind IR. We adopt the commonly used ResUNet [54], [56] in low-level vision as our backbone. It replaces the plain convolution layer in UNet [75] with residual block [76], and thus makes the gradient flow propagate much faster. Furthermore, in purpose of leveraging the estimated noise and kernel information by SNet and KNet, we concatenate their outputs with the corrupted image $y$ together (see Fig. 2), and then feed them into RNet to recover the high-resolution image. We empirically find that such a simple concatenated operation performs very well and stably in our inference framework.

It should be noted that this work does not aim to design more effective network architectures to surpass current SotA methods but to devise a probabilistic framework based on the deep variational inference for blind IR. Therefore, we simply employ the commonly used networks in low-level vision as our backbones. To better verify the generality of the proposed framework, we also test its effectiveness on other popular network architectures in Sec. 5.4.3.

5 EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of our proposed method on two typical IR tasks, namely image denoising and image super-resolution. We denote our Variational Image Restoration Network as VIRNet for notation convenience in the following presentation.

To optimize the network, we adopted the Adam [77] algorithm with a mini-batch size of 16 and other default settings of PyTorch [78]. The initial learning rate was set as $10^{-4}$ and decayed gradually using the cosine annealing strategy [79]. For computational stability, the gradient clipping strategy was also used during training. In the task of image denoising, we cropped small image patches with a size of $128 \times 128$ for training. The hyper-parameter $\varepsilon^2_0$ of Eq. (3) was set to be $10^{-6}$, and the window size $p$ of Eq. (5) was set to be 7. In the task of image super-resolution, the patch size during training was fixed as 96, 144, and 192 for scale factors 2, 3, and 4 respectively. The hyper-parameter $\varepsilon^2_0$ was set to be $10^{-3}$, while the window size $p$ was set as a larger value 11 than that in denoising, since image noise in super-resolution is usually assumed to be i.i.d. Gaussian. As for the shape parameter $\kappa_0$ of the kernel prior distribution in Eq. (11), we empirically set it as 50.

5.1 Image Denoising Experiments

5.1.1 Synthetic Non-I.I.D. Gaussian Noise Removal

To verify the effectiveness and robustness of VIRNet under non-i.i.d. noise configurations, we synthesized a large set of noisy/clean image pairs as training data. Similar to [20], a set of high-quality source images was firstly collected as clean ones, including 432 images from BSD500 [80], 400 images from the ImageNet [81] validation set and 4744 images from Waterloo Database [82]. We then randomly generated non-i.i.d. Gaussian noise as

$$n = n^1 \odot M, \quad n^1 \sim \mathcal{N}(n^1|0, I),$$

where $I$ is the identity matrix, and $M$ is a spatially vari-ant map with the same size as the source image. Finally, the noisy image was obtained by adding the generated noise $n$ to the source image. As for the testing images, two commonly-used datasets, i.e., BSD68 [80] and McMaster [83], were adopted to evaluate the performance of different methods. Note that we totally generated four kinds of $M$s as shown in Fig. 3. The first one (Fig. 3 (a)) was used for generating noisy images in training data, and the others (Fig. 3 (b1)-(d1)) for three groups of testing data (denoted as Cases 1-3). Under these settings, the noise in training data and testing data are evidently different, which is suitable to verify the generalization capability of VIRNet.

Comparison with the SotAs. We compare VIRNet with several current denoising methods, including two typical model-based methods NLM [3] and CBM3D [4], two deep self-supervised methods S2S [73] and Ne2Ne [74], three supervised learning-based methods, namely DnCNN [17], FFDNet [20], and DRUNet [54]. The PSNR and SSIM results of all comparison methods on three groups of testing data are listed in Table 1. We can easily see that: 1) the proposed VIRNet outperforms the other methods in all cases, indicating its superiority in handling these complicated
non-i.i.d. noise types; 2) on the whole, DL-based methods (including the self-supervised methods) evidently surpass classical model-based methods NLM and CBM3D, owning to the powerful non-linear fitting capability of DNNs; 3) FFDNet and DRUNet are both non-blind methods that rely on the pre-given noise level. In contrast, VIRNet is designed toward blind IR, and is thus able to simultaneously infer the noise distribution and remove the noise. Even so, VIRNet still achieves obvious performance improvements compared with FFDNet and DRUNet. This indicates the effectiveness of the Bayesian generative model and the variational inference framework.

Fig. 4 shows the visual results of different methods under testing cases 1-3 of Table 1. Note that we only display the best five DL-based methods due to page limitation. It can be easily seen that the comparison methods can remove most of the noises, but also often generate over-smooth and blurry recovery, especially in the heavy-noise areas. This can be explained by the fact that they do not consider spatial noise variations. To handle such non-i.i.d. noise, the proposed VIRNet elaborately considers the noise configurations and is thus capable of preserving more image details (e.g., edges, structures) than other methods.

Even though our VIRNet is designed and trained on the non-i.i.d. noise settings, it also performs well in additive white Gaussian noise (AWGN) removal tasks. Table 2 lists the average PSNR and SSIM results of different methods under three noise levels (i.e., $\sigma = 15, 25, 50$) of AWGN. In this part, we further add the method RNAN [21] for a more thorough evaluation. It is noteworthy that RNAN is separately trained on some specific noise levels for AWGN, and hence we can only compare with it on the noise level 50. It is easy to see that VIRNet obtains the best (8 out of 12 cases) or at least second-best (4 out of 12 cases) performance compared with these comparison methods. Combining the results in Table 1 and Table 2, it should be rational to say that the proposed VIRNet is more robust. Specifically, it is more flexible noise modeling essence.

In Table 3, we further list the comparison results on the number of model parameters and FLOPs with four DL-based methods. The FLOPs listed in this table are calculated based on images with a size of $512 \times 512$. It should be noted that, for the sake of fair comparisons, the self-supervised methods S2S [73] and Ne2Ne [74] are not reported in Table 3.
It can be easily observed that VIRNet exhibits a better compromise over current SotA methods RNAN [21] and DRUNet [46] when taking both the model parameters and FLOPs into consideration. The proposed VIRNet is thus expected with better practical applicability in real scenarios.

5.1.2 Real-world Noise Removal

In this part, we evaluate the performance of VIRNet on two widely used real-world benchmark datasets, namely DND [84] and SIDD [65]. DND consists of 50 high-resolution images with realistic noise from 50 scenes taken by 4 consumer cameras, but it does not provide any other noisy/clean image pairs as training data. SIDD is another real-world denoising benchmark, containing about 30,000 real noisy images captured by 5 cameras under 10 scenes. Different from DND, each noisy image in SIDD comes with an almost noise-free counterpart as ground truth, which is estimated by some statistical methods [65]. Further, SIDD also provides a small version dataset containing 320 image pairs, called SIDD-Medium, which is commonly used as training data in recent works [23], [24], [27]. To compare with them fairly, we also train VIRNet only based on the SIDD-Medium dataset. As for the metrics, we adopt PSNR and SSIM [85] calculated on the sRGB space to quantitatively evaluate different methods.


<table>
<thead>
<tr>
<th>Methods</th>
<th>SIDD</th>
<th>DND</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>SSIM</td>
<td>PSNR</td>
</tr>
<tr>
<td>CBM3D [4]</td>
<td>25.65</td>
<td>0.685</td>
</tr>
<tr>
<td>DnCNN [17]</td>
<td>23.66</td>
<td>0.583</td>
</tr>
<tr>
<td>CBDNet [22]</td>
<td>30.78</td>
<td>0.801</td>
</tr>
<tr>
<td>RIDNet [23]</td>
<td>38.71</td>
<td>0.951</td>
</tr>
<tr>
<td>AINDNet [86]</td>
<td>38.95</td>
<td>0.952</td>
</tr>
<tr>
<td>VDN [27]</td>
<td>39.23</td>
<td>0.955</td>
</tr>
<tr>
<td>SADNet [67]</td>
<td>39.46</td>
<td>0.957</td>
</tr>
<tr>
<td>DANet [24]</td>
<td>39.47</td>
<td>0.957</td>
</tr>
<tr>
<td>CycleISP [88]</td>
<td>39.52</td>
<td>0.957</td>
</tr>
<tr>
<td>MPRNet [25]</td>
<td>39.62</td>
<td>0.958</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td>39.63</td>
<td>0.958</td>
</tr>
<tr>
<td>PNGAN [89]</td>
<td>40.06</td>
<td>0.960</td>
</tr>
</tbody>
</table>

Note that PNGAN follows the same network architecture and the training strategy with MPRNet, and further improves the performance by simulating more training data. Thus we mark its results with gray color to denote unfair comparisons.
We compared VIRNet with several typical real-world denoising methods, including MPRNet [25], CycleISP [88], DANet [24], SADNet [87], VDN [27] and so on (see Table 4). To the best of our knowledge, current SotA method on these two benchmarks is PNGAN [89]. This work, however, mainly focuses on simulating the camera pipeline to generate large amount of image pairs as training data so as to further improve the performance, instead of designing more effective denoising algorithm. Its denoiser architecture and the training strategy completely follow MPRNet. Therefore we mainly compare with MPRNet in this work.

In order to comprehensively evaluate all competing methods, Table 4 lists the denoising performance, as well as the model profiles, including the number of network parameters, the FLOPs, and the feedforward running time of the denoisers. The FLOPs and running time are both counted on images with a size of \(512 \times 512\). From the perspective of denoising performance, the proposed VIRNet achieves a slight performance improvement compared with current SotA method MPRNet, indicating its effectiveness. However, VIRNet is of pronounced superiority in terms of running time, which more faithfully reflect the relative efficiency of our method. To intuitively compare the denoising results, we visualize two typical real examples in Fig. 5, which are consistent with the quantitative results in Table 4.

### 5.2 Image Super-resolution Experiments

In this section, we apply our proposed VIRNet to blind image super-resolution. Following [70], the DF2K dataset (containing 800 images from DIV2K [92] and 2650 images from Flickr2K [93]) was employed as our training data.

![Fig. 6. Seven Gaussian kernels used to generate the LR images in the synthetical super-resolution experiments.](image)

To be capable of quantitatively evaluating different methods, we first conducted some synthetic experiments on three commonly used datasets, including Set14 [91], CBSD68 [80], and DIV2K100 (the validation set of DIV2K [92]). To make a thorough comparison on various degradation, we considered seven representative and diverse kernels as shown in Fig. 6, including three isotropic Gaussian kernels with different kernel widths (i.e., 0.4s, 0.6s, and 0.8s) and four anisotropic Gaussian kernels, where \(s\) is the scale factor. In addition, three noise levels (i.e., 0, 2.55, and 7.65) are considered following [56]. As for the metrics, aside from the commonly used PSNR and SSIM [85], we also adopted LPIPS [95] to measure the perceptual similarity. Note that PSNR and SSIM are calculated on the Y channel of the YCbCr space, while LPIPS is calculated in the sRGB space.

We consider three categories of comparison methods: 1) classical Bicubic interpolation method; 2) five blind super-resolution methods, including HAN [90], IKC [44], DAN [57], DASR [70], and BSRNet [46]; 3) two non-blind methods, i.e., SRMD [89] and USRNet [56], which rely on the pre-given blur kernel and noise level as input. For these non-blind methods, we provided the ground truth blur kernel and noise level for them and denoted their results with the format of “GT+X” (e.g., GT+SRMD). In addition, for the methods of HAN, IKC and DAN, we firstly denoised the noisy low-resolution image using DnCNN [17] and then super-resolved it in the cases of \(\sigma = \{2.55, 7.65\}\), since these methods do not consider image noise during training.

Table 5 lists the comparison results of different methods under a scale factor of 4, and more results under scale factors of 2 and 3 are put into Appendix C. From Table 5, it can be seen that the proposed VIRNet achieves the best performance among blind methods in all cases. Especially, compared with the non-blind methods, VIRNet is still able to obtain slightly better or at least comparable results even though they make use of the ground truth information of blur kernel and noise level. This indicates the effectiveness of the proposed blind framework which is capable of handling the tasks of degradation estimation and image restoration simultaneously. Further, taking the model profiles into consideration, the superiority of VIRNet is more evident. Specifically, VIRNet has fewer number of parameters, fewer FLOPs, and faster speed than both the SotA blind method DASR [70] and the non-blind method USRNet [56].

In Fig. 7, we display the denoising results on three typical visual examples of Set14 with a scale factor of 4. Note that we only show the results of blind super-resolution methods for a fair comparison. It can be easily seen that the proposed VIRNet is able to recover more realistic and sharper results, which are evidently closer to the ground truth high-resolution images than other methods. The results of most comparison methods are relatively blurry and lose some image details. In the second example (the middle row), IKC and DAN lead to a relatively severe corruption of the original image color. That’s possibly caused by the inconsistency of their multiple iterations, since they both adopt the coarse-to-fine manner to gradually adjust the results. Due to the careful considerations of the degradation model, DASR and BSRNet also perform well compared with

---

4. Extracting the upper-left pixel for each \(p \times p\) patch.
TABLE 5
Quantitative comparisons of different methods under a scale factor of 4. The PSNR/SSIM/LPIPS values in this table are all averaged over the seven kernels as shown in Fig. 6. The results of the non-blind methods that rely on the pre-given ground-truth blur kernel and noise level are marked in gray color to denote unfair comparisons. Besides, some commonly considered model profiles, namely, the number of learnable parameters (in M), the FLOPs (in G), and the running time (in seconds), are also listed for more comprehensive comparison. Note that the FLOPs and running time are calculated in the case of super-resolving the low-resolution images with a size of 256 × 256.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Noise Level</th>
<th>Set14</th>
<th>CBSD68</th>
<th>DIV2K100</th>
<th></th>
<th></th>
<th># Param</th>
<th>FLOPs</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td></td>
<td>24.54</td>
<td>0.6352</td>
<td>0.5257</td>
<td>24.68</td>
<td>0.6144</td>
<td>0.6044</td>
<td>25.50</td>
<td>0.6/62</td>
</tr>
<tr>
<td>HAN [90]</td>
<td></td>
<td>25.36</td>
<td>0.6731</td>
<td>0.4693</td>
<td>25.35</td>
<td>0.6494</td>
<td>0.5477</td>
<td>26.30</td>
<td>0.7115</td>
</tr>
<tr>
<td>IKC [44]</td>
<td></td>
<td>27.24</td>
<td>0.7388</td>
<td>0.3534</td>
<td>26.70</td>
<td>0.7105</td>
<td>0.4253</td>
<td>28.00</td>
<td>0.7741</td>
</tr>
<tr>
<td>DAN [70]</td>
<td></td>
<td>27.49</td>
<td>0.7464</td>
<td>0.3442</td>
<td>26.96</td>
<td>0.7204</td>
<td>0.4053</td>
<td>28.31</td>
<td>0.7832</td>
</tr>
<tr>
<td>DASR [70]</td>
<td></td>
<td>27.74</td>
<td>0.7512</td>
<td>0.3314</td>
<td>27.12</td>
<td>0.7231</td>
<td>0.4026</td>
<td>28.32</td>
<td>0.7805</td>
</tr>
<tr>
<td>BSRNet [46]</td>
<td></td>
<td>26.84</td>
<td>0.7129</td>
<td>0.3819</td>
<td>26.38</td>
<td>0.6860</td>
<td>0.4540</td>
<td>27.49</td>
<td>0.7503</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td>0.10</td>
<td>27.89</td>
<td>0.7573</td>
<td>0.3165</td>
<td>27.27</td>
<td>0.7330</td>
<td>0.3868</td>
<td>28.60</td>
<td>0.7919</td>
</tr>
<tr>
<td>GT+SRMD [69]</td>
<td></td>
<td>27.83</td>
<td>0.7587</td>
<td>0.3180</td>
<td>27.12</td>
<td>0.7283</td>
<td>0.3962</td>
<td>28.44</td>
<td>0.7874</td>
</tr>
<tr>
<td>GT+USRNet [56]</td>
<td></td>
<td>27.90</td>
<td>0.7747</td>
<td>0.3818</td>
<td>26.88</td>
<td>0.7408</td>
<td>0.3988</td>
<td>28.64</td>
<td>0.8086</td>
</tr>
</tbody>
</table>

TABLE 6
The non-reference comparison results of different methods on the real-world dataset SRRealSet [46] under a scale factor of 4. The best and second best results are highlighted by bold and underline, respectively.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Methods</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NRQM↓</td>
<td>Bicubic</td>
<td>3.35</td>
<td>4.53</td>
<td>4.72</td>
<td>4.43</td>
<td>4.18</td>
<td>4.30</td>
<td>4.65</td>
<td></td>
</tr>
<tr>
<td>PI↓</td>
<td></td>
<td>6.36</td>
<td>5.35</td>
<td>5.33</td>
<td>5.34</td>
<td>5.73</td>
<td>5.90</td>
<td>5.39</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 7
The PSNR results of FFDNet [20] with the groundtruth noise level and with that estimated by the proposed VIRNet on CBSD68 [80] under the i.i.d. Gaussian noise setting.

<table>
<thead>
<tr>
<th>No. Levels</th>
<th>Methods</th>
<th>Noise Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ = 15</td>
<td>FFDNetGT</td>
<td>33.87</td>
</tr>
<tr>
<td>σ = 30</td>
<td>FFDNetVIR</td>
<td>33.87</td>
</tr>
</tbody>
</table>

Fig. 7. Visualized super-resolution results of different methods on a synthetic example of Set14 [91] under a scale factor of 4. Specifically, the noise level is 2.55, and the blur kernel is shown on the upper-right corner of the LR image.

5.2.2 Results on Real Data
In this part, we further justify the effectiveness of the proposed VIRNet on the real-world dataset RealSRSet [46]. It...

other methods. However, VIRNet still evidently surpasses them in terms of the quantitative and qualitative results. This further substantiates the effectiveness of the proposed variational framework.
contains 20 real images that are commonly used in previous literature [20], [97]–[99] or downloaded from the internet. Since the underlying high-resolution images for them are not available, we thus mainly evaluate different methods by visual comparisons. Fig. 8 displays three typical super-resolution examples with a scale factor of 4. In the first (top row) and second (middle row) examples, the LR images both contain some image noises, which makes the super-resolution goal more challenging. The methods of Bicubic, HAN, IKC, and DAN all fail to deal with such cases and produce some artifacts in the areas with image noises. As for DASR and BSRNet, they, unfortunately, erase the high-frequency image details when removing the image noises. One can easily observe that the proposed VIRNet makes a good trade-off between preserving the image details and removing the image noises. In the third (bottom row) example, the results of IKC and VIRNet are more natural and realistic than others that are all blurry to different extents. These results verify the stable and consistently well performance of VIRNet in the real-world super-resolution task.

In Table 6, we adopt two non-reference metrics (i.e., NRQM [100] and PI [101]) to further quantitatively evaluate different methods. It can be seen that the proposed VIRNet achieves the second best results in terms of both metrics, only slightly worse than IKC, which indicates the effectiveness of our method. Combining its better visual performance as shown in Fig. 8, especially compared with those obtained by IKC, the relative superiority of the proposed method can still be validated.

### 5.2.3 Discussion on Blur Kernel

How to properly and generally set the degradation model, especially the blur kernel, is still an open yet challenging research topic in the field of image super-resolution [46], [102]. In this work, we adopt the anisotropic Gaussian assumption for the blur kernel following most of current related literature [44], [57], [63], [68]–[70]. Even though such general kernel hypothesis is sufficient in most of scenarios, it still possibly leads to some unsatisfactory results in some cases. For instance, Fig. 9 shows one typical failed exam-
In this example, the LR image contains some obvious “ringing artifacts” that look like bands or ghosts near edges, which are usually produced by the sharpening algorithm or image compression. The Gaussian kernel-based methods (i.e., IKC [44], DAN [57], DASR [70], and the proposed VIRNet) all cannot finely resolve this image, while BSRNet [46] performs well in this case, mainly because it integrates multiple complicated kernel settings. Therefore, it is necessary to exploit a more rational and general kernel (or degradation) modeling method for image super-resolution, and we leave this to future work.

5.3 Degradation Estimation Experiments

In this subsection, we empirically verify the effectiveness of our method in the task of degradation estimation, including noise estimation and kernel estimation.

5.3.1 Noise Estimation

Different from most of the current IR methods, the pixel-wise non-i.i.d. Gaussian assumption is adopted to fit the noise distribution in our method. Next, we analyze the performance of our method with such an assumption under several common noise types in IR tasks:

I.I.D. Gaussian Noise. Even though VIRNet is designed based on non-i.i.d. Gaussian noise assumption, it can be generalized well to the i.i.d. Gaussian noise as shown in Table 2. To further quantitatively illustrate this point, we take the estimated noise level by our method as the input of FFDNet [20], which is a typical non-blind i.i.d. Gaussian denoising method that relies on the pre-known noise level. The proposed method performs well in this case, mainly because it integrates multiple complicated kernel settings. Therefore, it is necessary to exploit a more rational and general kernel (or degradation) modeling method for image super-resolution, and we leave this to future work.

Signal-dependent Noise. The challenge in real-world image denoising is mainly attributed to the signal-dependentness of the image noise. Fig. 10 shows one typical real-world noisy example coming from SIDD [65] dataset and the corresponding variance maps predicted by VIRNet. Note that the variance map has been enlarged several times for easy visualization. It is easy to see that the estimated noise variance map depicts strong relevance to the pixel illumination, implying that the proposed VIRNet is able to finely approximate the signal-dependent real noise.

5.3.2 Kernel Estimation

As is well known, kernel estimation plays an important role in blind image super-resolution [103]. To evaluate the effectiveness of VIRNet in this subtask, we compare three recent kernel estimation methods specifically designed for super-resolution, including KernelGan [96], DIPFKP [63], and BSRDM [64]. Since these three methods are all relatively time-consuming, we randomly select 20 images from the validation set of DIV2K [92] (denoted as DIV2K20) as testing data. The LR images are synthesized using the last four anisotropic Gaussian kernels in Fig. 6 under a scale factor of 4, and the noise level is set as 2.55.

As for the evaluation, we use two ways to compare the performance of different methods. Firstly, the MSE between the estimated kernel and the ground truth kernel is an intuitive metric that directly reflects the accuracy of the estimated kernel. The detailed comparison results are listed in Table 9. Secondly, we apply the estimated blur kernels in a non-blind super-resolution method USRNNet [56] and then compare the recovered HR images in terms of PSNR, SSIM [85], and LPIPS [95]. The comparison results on these three metrics are listed in Table 10. From both tables, one can easily observe that the proposed VIRNet exhibits evident superiority over other competing methods.

5.4 Ablation Study

5.4.1 Posteriori Factorization

When designing the variational inference algorithm in Sec. 4, we factorize the variational distribution \( q(z, \sigma^2, A | y) \) into a conditional format of Eq. (12), which fundamentally induces the cascaded inference framework in Fig. 2. In fact, different factorized assumptions on \( q(z, \sigma^2, A | y) \) will lead to different designs on the inference framework. For example, the following unconditional factorization

\[
q(z, \sigma^2, A | y) = q(z | y)q(\sigma^2 | y)q(A | y),
\]

will induce a parallel inference architecture. Specifically, the three sub-networks, namely SNet, KNet, and RNet, will feed-forward independently in such a parallel framework, but they can interact during backpropagation through ELBO. Please refer to our previous conference version [27] for a thorough overview on this point.
To validate the superiority of the conditional form of Eq. (12), we consider different posteriori factorization and empirically compare their performance on the task of image super-resolution, since it involves more general degradation than image denoising. Table 11 lists the average comparison results on DIV2K100 with a scale factor of 4. As compared with Baseline1, it can be seen that VIRNet achieves evident performance gain, which indicates that the degradation information (i.e., the noise level and the blur kernel) can facilitate the image restoration task. In fact, such a conditional factorization in VIRNet is consistent with the classical model-based methods that decompose the blind IR in two subproblems, namely degradation estimation and image restoration. The superiority of Baseline3 over Baseline2 demonstrates that the kernel information can bring up more marginal performance improvement than the noise level, complying with the conclusion in [103]. However, the performance gain of VIRNet over Baseline3 on LPIPS substantiates that conditioning on the noise level can further improve the perceptual quality of the recovered images.

### 5.4.3 Restoration Backbones

Our proposed method introduces a unified Bayesian framework to handle the task of blind IR, which does not rely on specific designs on the network architecture. By taking the limit as $\varepsilon_0 \to 0$, the framework then degenerates into the widely used MSE loss function. To comprehensively validate the superiority of our method beyond MSE, we consider four typically popular backbones, namely DnCNN [17], UNet [75], ResUNet [54], and SwinIR [104], for the restoration network RNet in our method. We train these four backbones with MSE under the same settings to our method as baselines. Quantitative comparison on the task of real-world image denoising is itemized in Table 14. It can be seen that the proposed method brings up evident performance gain as compared with the MSE baseline, particularly for UNet and SwinIR, indicating the superiority of the proposed framework.

### 6 Conclusion

In this paper, we have proposed a novel deep variational network for blind IR, which aims to finely integrate the merits of both classical model-based methods and recent DL-based methods. On one hand, we have constructed a Bayesian generative model for blind image denoising and image super-resolution, by carefully considering the image degradation from the perspectives of image noise and blur.
kernel. On the other hand, a variational inference algorithm has been elaborately designed to solve the proposed model, in which the posterior distribution is all parameterized by DNNs to increase the non-linear fitting capability. Most notably, this variational algorithm induces a unified framework to simultaneously deal with the tasks of degradation estimation and image restoration. Extensive experiments have also been conducted to demonstrate the superiority of our method on image denoising and super-resolution. In the future, we will make further efforts to extend our method to deal with more complicated and general image degradation. Besides, we will make efforts to explore more general kernel types beyond the Gaussian for SR and other image restoration tasks under our proposed statistical framework.

ACKNOWLEDGMENTS

This study is supported by the National Key R&D Program of China (2020YFA0713900), the China NSFC projects under contracts 62076196, 62331028, 12226004.

REFERENCES


Hongwei Yong received his Ph.D. degree from the Department of Computing at the Hong Kong Polytechnic University in 2022 and received his MSc. degree and B.Sc degree from Xi’an Jiaotong University, China, in 2016 and 2013, respectively. He is currently a research assistant professor in the Department of Computing, at the Hong Kong Polytechnic University (PolyU). His research interests focus on the research areas of computer vision and deep learning optimization. He has published a lot of research works in top journals and conferences including TPAMI, TIP, TNNLS, CVPR, ECCV, NeurIPS, etc. He is also a reviewer of some prestigious journals and conferences in my research areas, such as TPAMI, TIP, CVPR, AAAI, ICLR, NeurIPS, etc.

Qian Zhao (M’23) received the BSc and PhD degrees from Xi’an Jiaotong University, Xi’an, China, in 2009 and 2015, respectively. He was a visiting scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2013 to 2014. He is currently an Associate Professor with the School of Mathematics and Statistics, Xi’an Jiaotong University. His current research interests include low-rank matrix/tensor analysis, Bayesian modeling and meta learning.

Lei Zhang joined the Department of Computing, The Hong Kong Polytechnic University, as an Assistant Professor in 2006. Since July 2017, he has been a Chair Professor in the same department. His research interests include Computer Vision, Pattern Recognition, Image and Video Analysis, Deep Learning, etc. Prof. Zhang has published more than 200 papers in those areas. As of 2023, his publications have been cited more than 90,000 times in literature. Prof. Zhang is a Senior Associate Editor of IEEE Trans. on Image Processing, and is/was an Associate Editor of IEEE Trans. on Pattern Analysis and Machine Intelligence, SIAM Journal of Imaging Sciences, IEEE Trans. on CVST, and Image and Vision Computing, etc. He was listed as a “Clarivate Analytics Highly Cited Researcher” consecutively from 2015 to 2022. He is an IEEE Fellow. More information can be found in his homepage http://www4.comp.polyu.edu.hk/~cslzhang/.

Deyu Meng received the B.Sc., M.Sc., and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2012 to 2014. He is currently a professor with School of Mathematics and Statistics, Xi’an Jiaotong University, and adjunct professor with Faculty of Information Technology, The Macau University of Science and Technology. His current research interests include model-based deep learning, variational networks, and meta-learning.

Kwan-Yee K. Wong (Senior Member, IEEE) received the B.Eng. degree (Hons.) in computer engineering from The Chinese University of Hong Kong in 1998, and the M.Phil. and Ph.D. degrees in computer vision (information engineering) from the University of Cambridge in 2000 and 2001, respectively. Since 2001, he has been with the Department of Computer Science at The University of Hong Kong, where he is currently an Associate Professor. His research interests include computer vision and machine intelligence. He is currently an Editorial Board Member of International Journal of Computer Vision (IJCV).
A. Preliminaries on Inverse Gamma Distribution

The inverse Gamma distribution is a continuous statistical distribution on the positive real line, whose probability density function is defined as follows:

\[
f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left( \frac{1}{x} \right)^{\alpha+1} \exp \left( -\frac{\beta}{x} \right),
\]

where \( \Gamma(\cdot) \) denotes the Gamma function, \( \alpha \) and \( \beta \) are two parameters, called as “shape” and “scale”, respectively. For the inverse Gamma distribution, its mode is \( \frac{\beta}{\alpha + 1} \).

In Bayesian statistics, the inverse Gamma distribution is usually used as the conjugate prior for the unknown variance of a Gaussian distribution. In this study, we also follow this principle and set the prior for the noise variance as inverse Gamma distribution in Eq. (4) of the main text, i.e.,

\[
\sigma_i^2 \sim \text{IG} \left( \sigma_i^2 \mid \alpha_0 - 1, \alpha_0 \xi_i \right), \quad i = 1, 2, \ldots, d.
\]

With such settings for the shape and scale parameters, its mode just right equals \( \xi_i \) that is an empirical estimation for the noise variance, namely,

\[
\text{Mode}[\sigma_i^2] = \frac{\alpha_0 \xi_i}{(\alpha_0 - 1) + 1} = \xi_i.
\]

Furthermore, the shape parameter \( \alpha_0 \) controls the strength of such prior constraint. Specifically, the larger \( \alpha_0 \) is, the more accurate the estimated variance \( \xi_i \) is. This can be intuitively explained by Fig. 11.

Similarly, for the blur kernel with covariance matrix \( \Sigma = \begin{bmatrix} \lambda_1^2 & \lambda_1 \lambda_2 \rho \\ \lambda_1 \lambda_2 \rho & \lambda_2^2 \end{bmatrix} \), we also impose inverse Gamma distribution for \( \lambda_1^2 \) and \( \lambda_2^2 \) with shape parameter \( \kappa_0 \). Empirically, we find our method performs well and stably by setting \( \kappa_0 = 50 \) throughout all the experiments.

![Fig. 11. The probability density function of the inverse Gamma distribution IG(· | \alpha_0 - 1, \alpha_0 * \xi_i) under different values of \alpha_0 when \xi_i = 0.001.](image-url)
B. Details on Calculating the Evidence Lower Bound

In this section, we present the detailed calculations of some equations in the manuscript:

- Calculation details of Eq. (16):

\[
\log p(y) = \int q(z, \sigma^2, A \mid y) \log p(y \mid z, \sigma^2, A) \, dz \, d\sigma \, dA
\]

\[
= \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(y \mid z, \sigma^2, A) p(z) p(\sigma^2) p(A)}{p(z, \sigma^2, A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
= \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(y \mid z, \sigma^2, A) p(z) p(\sigma^2) p(A)}{q(z, \sigma^2, A \mid y)} + \frac{q(z, \sigma^2, A \mid y)}{p(z, \sigma^2, A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
= \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(y \mid z, \sigma^2, A) p(z) p(\sigma^2) p(A)}{q(z, \sigma^2, A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
+ \int q(z, \sigma^2, A \mid y) \log \left[ \frac{q(z, \sigma^2, A \mid y)}{p(z, \sigma^2, A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
= \mathcal{E}_{q(z, \sigma^2, A \mid y)} \left[ \log p(y \mid z, \sigma^2, A) p(z) p(\sigma^2) p(A) - \log q(z, \sigma^2, A \mid y) \right]
\]

\[
+ KL \left[ q(z, \sigma^2, A \mid y) \| p(z, \sigma^2, A \mid y) \right].
\]

- Calculation details of Eq. (18):

\[
\mathcal{L}(z, \sigma^2, A \mid y) = \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(y \mid z, \sigma^2, A) p(z) p(\sigma^2) p(A)}{q(z, \sigma^2, A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
= \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(y \mid z, \sigma^2, A) p(z) p(\sigma^2) p(A)}{q(z, \sigma^2, A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
+ \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(z)}{q(z \mid \sigma^2, A, y) q(\sigma^2 \mid y) q(A \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
+ \int q(z, \sigma^2, A \mid y) \log \left[ \frac{p(\sigma^2)}{q(\sigma^2 \mid y)} \right] \, dz \, d\sigma \, dA
\]

\[
= \mathcal{E}_{q(z, \sigma^2, A \mid y)} \left[ \log p(y \mid z, \sigma^2, A) \right]
\]

\[
- \mathcal{E}_{q(\sigma^2 \mid A \mid y)} \left[ KL \left[ q(z \mid \sigma^2, A, y) \| p(z) \right] \right]
\]

\[
- KL \left[ q(\sigma^2 \mid y) \| p(\sigma^2) \right]
\]

\[
- KL \left[ q(A \mid y) \| p(A) \right],
\]

where \(q(\sigma^2, A \mid y) = q(\sigma^2 \mid y) q(A \mid y)\).

- Calculation details of Eq. (20):

\[
\mathcal{E}_{q(z, \sigma^2, A \mid y)} \left[ \log p(y \mid z, \sigma^2, A) \right] = \int q(z, \sigma^2, A \mid y) \log p(y \mid z, \sigma^2, A) \, dz \, d\sigma \, dA
\]

\[
= \int q(z \mid y, \sigma^2, A) q(\sigma^2 \mid y) q(A \mid y) \log p(y \mid z, \sigma^2, A) \, dz \, d\sigma \, dA
\]

\[
\approx \frac{1}{L} \sum_{j=1}^{L} \log p(y^{(j)} \mid z^{(j)}, \sigma^{(j)^2}, A^{(j)})
\]

\[
\L(20)
\]
\[
\frac{1}{L} \sum_{j} \sum_{i} \log p(y_i | \tilde{z}^{(j)}, \tilde{\sigma}^{(j)^2}, \tilde{\Lambda}^{(j)^2})
\]
\[
= \frac{1}{L} \sum_{j} \sum_{i} \left\{ -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \tilde{\sigma}^{(j)^2} - \frac{\left( y_i - \left( \frac{\| \tilde{z}^{(j)} \|_2 \| \tilde{k}^{(j)} \|_2 \| \tilde{j}^{(j)} \|_2 \right)}{2\tilde{\sigma}^{(j)^2}} \right)^2}{2\tilde{\sigma}^{(j)^2}} \right\},
\]
(32)

where
\[
\tilde{z}^{(j)} = \mu + \varepsilon_0 \epsilon^{(j)}, \quad \epsilon^{(j)} \sim \mathcal{N}(\epsilon | 0, I_d),
\]
(33)
\[
\tilde{k}^{(j)} = g \left( \tilde{\Lambda}^{(j)} \right), \quad \tilde{\Lambda}^{(j)} \sim q(\Lambda | y),
\]
(34)
\[
\tilde{\sigma}^{(j)^2} \sim q(\sigma^2 | y).
\]
(35)

After setting \( L = 1 \), we can obtain the Eq. (20) of the manuscript.

- The KL divergence of Gaussian and inverse Gamma distributions in Eqs. (21)-(23) of the manuscript can be found in \([105]\) and \([106]\), respectively.

C. Additional Experimental Results on Image Supre-resolution

In Table 5 of the manuscript, we list the quantitative comparison results of different methods under scale factor 4 for the task of image super-resolution. Here we additionally present the results under scales 2 and 3 in Table 15 and Table 16, respectively, for more comprehensive comparison. Obviously, compared with the results on scale factor 4 of the manuscript, the proposed method exhibits more evident superiorities under scale factors 2 and 3, which further demonstrates the effectiveness of our method.
TABLE 15
Quantitative comparisons of different methods under scale factor 2. The PSNR/SSIM/LPIPS values in this table are all averaged over the seven kernels as shown in Fig. 5 of the manuscript. The results of the non-blind methods that rely on the pre-given ground truth blur kernel and noise level are marked in gray color to denote unfair comparisons.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Noise Level</th>
<th>Set14</th>
<th>CBSD68</th>
<th>DIV2K100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
<td>LPIPS</td>
</tr>
<tr>
<td>Bicubic</td>
<td></td>
<td>28.04</td>
<td>0.7940</td>
<td>0.3096</td>
</tr>
<tr>
<td>DnCNN [17]+HAN [90]</td>
<td></td>
<td>29.25</td>
<td>0.8301</td>
<td>0.2643</td>
</tr>
<tr>
<td>IKE [44]</td>
<td></td>
<td>31.61</td>
<td>0.8868</td>
<td>0.1700</td>
</tr>
<tr>
<td>DAN [57]</td>
<td></td>
<td>32.67</td>
<td>0.9048</td>
<td>0.1310</td>
</tr>
<tr>
<td>DASR [70]</td>
<td></td>
<td>32.19</td>
<td>0.8957</td>
<td>0.1586</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td></td>
<td>32.70</td>
<td>0.9047</td>
<td>0.1194</td>
</tr>
<tr>
<td>GT+USRNet [56]</td>
<td></td>
<td>32.85</td>
<td>0.9041</td>
<td>0.1348</td>
</tr>
<tr>
<td>GT+SRMD [69]</td>
<td></td>
<td>31.52</td>
<td>0.9066</td>
<td>0.1463</td>
</tr>
<tr>
<td>Bicubic</td>
<td></td>
<td>27.97</td>
<td>0.7872</td>
<td>0.3084</td>
</tr>
<tr>
<td>DnCNN [17]+HAN [90]</td>
<td></td>
<td>28.95</td>
<td>0.8096</td>
<td>0.2565</td>
</tr>
<tr>
<td>DnCNN [17]+IKC [44]</td>
<td></td>
<td>29.35</td>
<td>0.8241</td>
<td>0.2459</td>
</tr>
<tr>
<td>DnCNN [17]+DAN [57]</td>
<td></td>
<td>30.01</td>
<td>0.8391</td>
<td>0.1946</td>
</tr>
<tr>
<td>DASR [70]</td>
<td></td>
<td>31.02</td>
<td>0.8620</td>
<td>0.2106</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td></td>
<td>31.48</td>
<td>0.8695</td>
<td>0.1764</td>
</tr>
<tr>
<td>GT+USRNet [56]</td>
<td></td>
<td>31.62</td>
<td>0.8713</td>
<td>0.1795</td>
</tr>
<tr>
<td>GT+SRMD [69]</td>
<td></td>
<td>31.08</td>
<td>0.8798</td>
<td>0.1933</td>
</tr>
</tbody>
</table>

TABLE 16
Quantitative comparisons of different methods under scale factor 3. The PSNR/SSIM/LPIPS values in this table are all averaged over the seven kernels as shown in Fig. 5 of the manuscript. The results of the non-blind methods that rely on the pre-given ground truth blur kernel and noise level are marked in gray color to denote unfair comparisons.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Noise Level</th>
<th>Set14</th>
<th>CBSD68</th>
<th>DIV2K100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
<td>LPIPS</td>
</tr>
<tr>
<td>Bicubic</td>
<td></td>
<td>25.89</td>
<td>0.6988</td>
<td>0.4459</td>
</tr>
<tr>
<td>HAN [90]</td>
<td></td>
<td>26.82</td>
<td>0.7374</td>
<td>0.3884</td>
</tr>
<tr>
<td>IKE [44]</td>
<td></td>
<td>27.76</td>
<td>0.7966</td>
<td>0.2805</td>
</tr>
<tr>
<td>DAN [57]</td>
<td></td>
<td>29.08</td>
<td>0.8108</td>
<td>0.2699</td>
</tr>
<tr>
<td>DASR [70]</td>
<td></td>
<td>29.26</td>
<td>0.8147</td>
<td>0.2740</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td></td>
<td>29.44</td>
<td>0.8190</td>
<td>0.2518</td>
</tr>
<tr>
<td>GT+USRNet [56]</td>
<td></td>
<td>29.75</td>
<td>0.8243</td>
<td>0.2488</td>
</tr>
<tr>
<td>GT+SRMD [69]</td>
<td></td>
<td>29.28</td>
<td>0.8332</td>
<td>0.2426</td>
</tr>
<tr>
<td>Bicubic</td>
<td></td>
<td>25.84</td>
<td>0.6936</td>
<td>0.4637</td>
</tr>
<tr>
<td>DnCNN [17]+HAN [90]</td>
<td></td>
<td>26.65</td>
<td>0.7207</td>
<td>0.4064</td>
</tr>
<tr>
<td>DnCNN [17]+IKC [44]</td>
<td></td>
<td>27.00</td>
<td>0.7362</td>
<td>0.3661</td>
</tr>
<tr>
<td>DnCNN [17]+DAN [57]</td>
<td></td>
<td>26.85</td>
<td>0.7292</td>
<td>0.3932</td>
</tr>
<tr>
<td>DASR [70]</td>
<td></td>
<td>28.54</td>
<td>0.7819</td>
<td>0.3029</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td></td>
<td>28.66</td>
<td>0.7843</td>
<td>0.2914</td>
</tr>
<tr>
<td>GT+USRNet [56]</td>
<td></td>
<td>28.90</td>
<td>0.7893</td>
<td>0.2809</td>
</tr>
<tr>
<td>GT+SRMD [69]</td>
<td></td>
<td>28.84</td>
<td>0.8079</td>
<td>0.2683</td>
</tr>
<tr>
<td>Bicubic</td>
<td></td>
<td>25.88</td>
<td>0.6957</td>
<td>0.5777</td>
</tr>
<tr>
<td>DnCNN [17]+HAN [90]</td>
<td></td>
<td>25.62</td>
<td>0.6239</td>
<td>0.5759</td>
</tr>
<tr>
<td>DnCNN [17]+IKC [44]</td>
<td></td>
<td>26.51</td>
<td>0.7116</td>
<td>0.3652</td>
</tr>
<tr>
<td>DnCNN [17]+DAN [57]</td>
<td></td>
<td>26.25</td>
<td>0.6746</td>
<td>0.4002</td>
</tr>
<tr>
<td>DASR [70]</td>
<td></td>
<td>27.62</td>
<td>0.7415</td>
<td>0.3397</td>
</tr>
<tr>
<td>VIRNet (Ours)</td>
<td></td>
<td>27.65</td>
<td>0.7431</td>
<td>0.3309</td>
</tr>
<tr>
<td>GT+USRNet [56]</td>
<td></td>
<td>27.75</td>
<td>0.7449</td>
<td>0.3337</td>
</tr>
<tr>
<td>GT+SRMD [69]</td>
<td></td>
<td>28.24</td>
<td>0.7690</td>
<td>0.3010</td>
</tr>
</tbody>
</table>