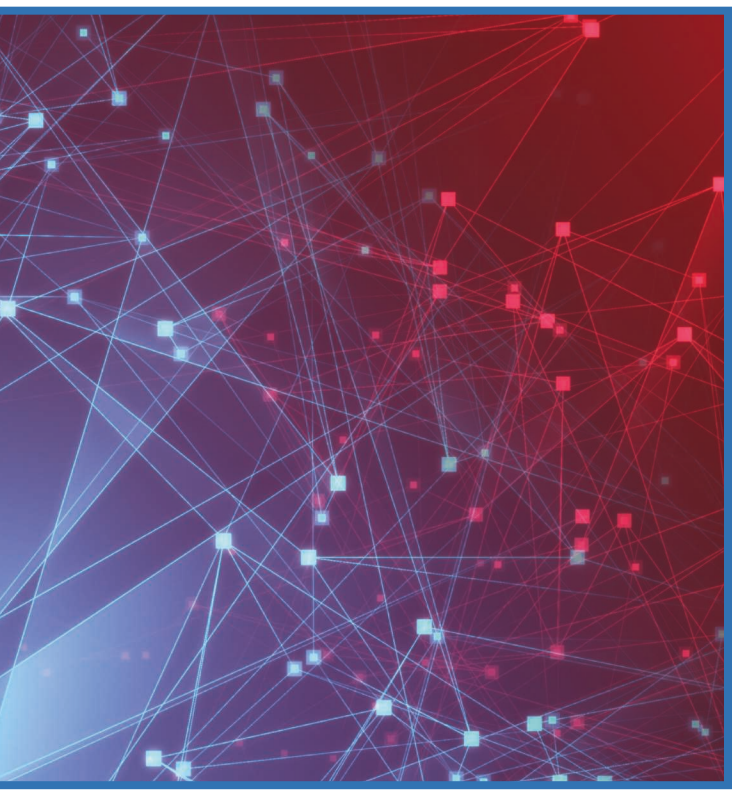
Weisheng Dong, Jinjian Wu, Leida Li, Guangming Shi, and Xin Li

# Bayesian Deep Learning for Image Reconstruction

*From structured sparsity to uncertainty estimation*



©SHUTTERSTOCK.COM/PAPAPIG

Conventional wisdom in model-based computational imaging incorporates physics-based imaging models, noise characteristics, and image priors into a unified Bayesian framework. Rapid advances in deep learning have inspired a new generation of data-driven computational imaging systems with performances even better than those of their model-based counterparts. However, the design of learning-based algorithms for computational imaging often lacks transparency, making it difficult to optimize the entire imaging system in a complete manner.

In this tutorial, we review the latest advances in deep learning that combine the strengths of model-based and learning-based approaches. By unfolding iterative optimization into a deep neural network implementation, we can sing an old folk song to a fast new tune. The explicit estimation of the uncertainty associated with the estimates allows us to construct a new class of uncertainty-driven loss (UDL) functions for deep unfolded networks. Using superresolution and depth imaging as examples, we demonstrate that the combination of deep neural networks and uncertainty modeling leads to the so-called "Bayesian deep learning" (BDL). Under the framework of BDL, we achieve a principled approach to model and estimate uncertainty for deep learning-based image reconstruction.

## Introduction

One of the enabling technologies in computational imaging is image reconstruction, which deals with the reconstruction of high-quality images from low-quality observations. Rapid advances in image reconstruction have evolved from model-based approaches [7], [8] to learning-based approaches [6], [27] in the past decade. In model-based image reconstruction, sparse coding has evolved from early local sparsity models (e.g., wavelet-based) to structured sparsity models (e.g., Block-matching and 3D filtering algorithm and low rank [5]). In learning-based image reconstruction, we have witnessed the impact of deep convolutional neural network (DCNN)-based image denoising [28]; nonlocal recurrent networks [16]; and, more recently, deep denoising priors for plug-and-play image reconstruction [6], [27]. Despite their outstanding performance,

*Digital Object Identifier 10.1109/MSP.2022.3176421*
*Date of current version: 29 December 2022*

the interpretability of learning-based methods is often not as transparent as that of model-based methods. Similar to the potential mismatch between a model and data, the generalization property of CNN-based image reconstruction remains poorly understood.

The motivation behind this article is mainly twofold. On the one hand, instead of mathematically constructing regularization functionals, data-driven surrogate models have been developed to incorporate domain-specific knowledge contained in physics-driven models [1]. Model-driven deep learning architectures, including deep unfolding networks (DUNs) [9] and plug-and-play image reconstruction [27], have found successful applications in various inverse problems, from medical imaging [11], [21] to image reconstruction [6], [27]. Several celebrated convex optimization algorithms, such as the iterative soft thresholding algorithm (ISTA) and the alternating-direction multiplier method (ADMM), have been unfolded into the corresponding neural network implementations with improved interpretability, namely, ISTA-net [26] and ADMM-net [22].

**Rapid advances in deep learning have inspired a new generation of data-driven computational imaging systems with performances even better than those of their model-based counterparts.**

On the other hand, uncertainty modeling has started to attract attention from the deep learning community in recent years [12], as shown in Figure 1. In both paradigms of model-based and learning-based image reconstruction, uncertainty often arises from either insufficient training data (e.g., due to cost constraints) or anomalous samples in the testing data (e.g., due to noise contamination). Bayesian statistics has evolved into a powerful framework for addressing uncertainty-related issues in a principled manner. In Bayesian modeling, there are two types of uncertainty to consider in image reconstruction: aleatoric uncertainty, which captures the noise inherent in observations, and epistemic uncertainty, which accounts for the uncertainty in the model, regardless of whether it is constructed mathematically or driven by data [12]. The latter can be explained away given enough data, at least in theory. However, it remains unclear how much data will be enough in practical situations; moreover, for the class of regression problems, such as blind image reconstruction, we have to deal with both uncertainties due to the lack of a priori information about either the unknown image or the real-world degradation process.

For the first time, BDL offers a principled framework for 1) leveraging the rich literature of regularized image reconstruction to construct a deep network with better interpretability

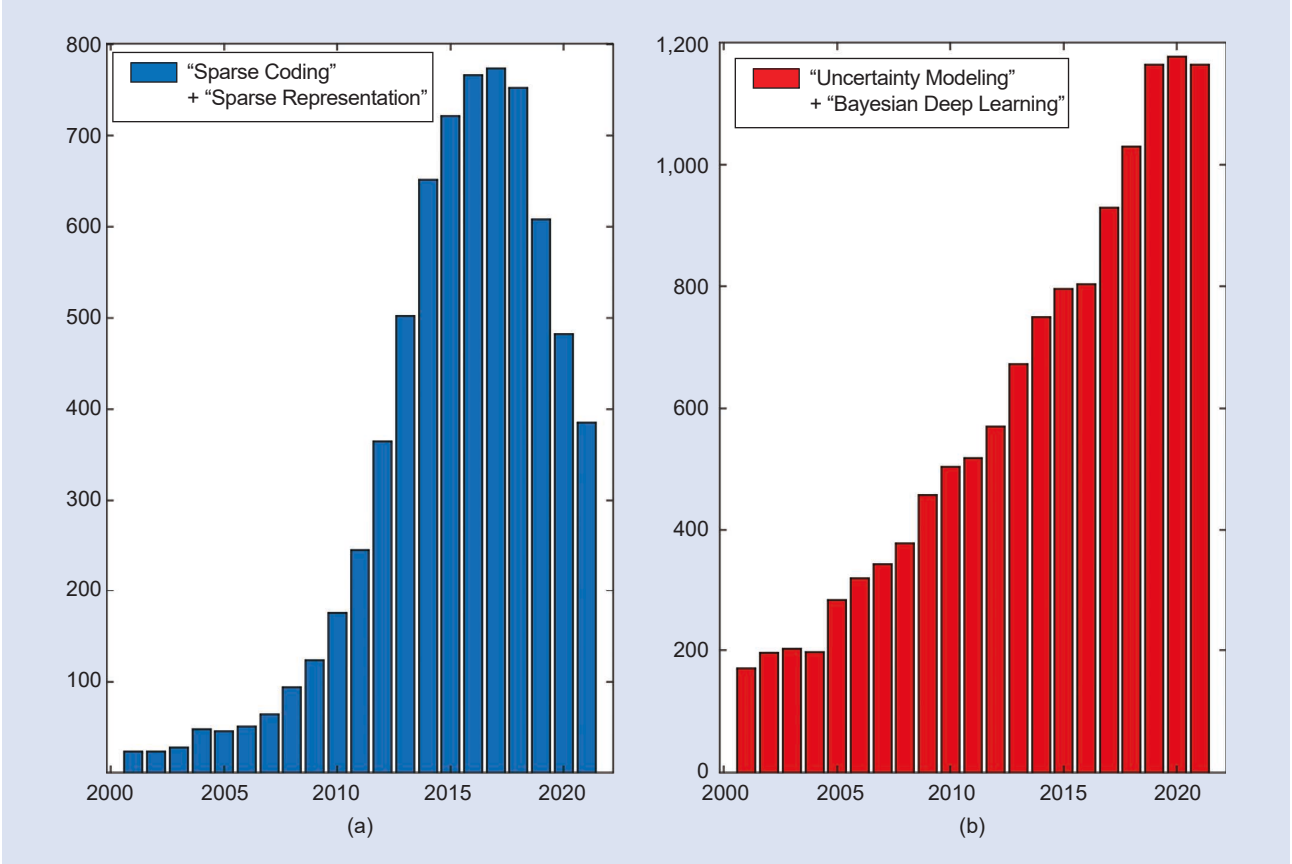



**FIGURE 1.** The number of publications on sparse coding and uncertainty modeling, searched from the Web of Science database with the title keywords (a) "sparse coding" (or "sparse representation") and (b) "uncertainty modeling" (or "Bayesian deep learning").

by deep unfolding and 2) improving the performance of CNN-based image reconstruction by taking uncertainties into account. The new insights brought about by this tutorial are summarized as follows:

- We conduct a systematic review of existing DUNs to bridge model-based and learning-based computational imaging. DUN offers a unified framework for translating various sparsity models into deep neural network implementations.
- We present a new framework for deep uncertainty-aware learning (DUAL) for image reconstruction capable of uncertainty modeling. In particular, we show how uncertainty modeling can lead to a novel design of UDL functions in image reconstruction.
- By combining DUN with DUAL, we advocate a novel approach to developing transparent learning-based solutions to image reconstruction problems by the end-to-end optimization of network parameters. Such a tuning-free property is desirable for task-specific optimization in various practical computational imaging applications.

**Bayesian statistics has evolved into a powerful framework for addressing uncertainty-related issues in a principled manner.**

## Past: From structured sparsity to DUNs

From 2000 to the present, we can divide the historical development of computational imaging techniques into two periods: sparsity based (2000–2016) and deep learning (2012–present). In this section, we first review the existing work on Bayesian formulations of sparse coding and their applications in image reconstruction. We then discuss some recent work on deep learning-based computational imaging.

### *Model-based image reconstruction via sparse coding*

We start from the Bayesian formulation of sparse coding. Then we will take into account the uncertainty related to the mean and variance, leading to extensions into nonlocal centralized and simultaneous sparse coding, respectively. The key idea of sparse coding is to decompose a signal $\boldsymbol{x} \in R^n$ ($n$ is the size of an image patch) into the linear combination of basis vectors (as well as dictionary elements) $\boldsymbol{D\alpha}$, where $\boldsymbol{D} \in R^{n \times K}$, $n \le K$ is the dictionary. The coefficients $\boldsymbol{\alpha} \in R^K$ satisfy some sparsity constraint, e.g., since $l_0$ optimization is computationally prohibitive to solve due to its nonconvexity, one can consider the $l_1$ counterpart as the surrogate function:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \| \boldsymbol{x} - \boldsymbol{D\alpha} \|_2^2 + \lambda \| \boldsymbol{\alpha} \|_1 . \tag{1}$$

In addition to convexity, solving the $l_1$-norm minimization problem is mathematically equivalent to the maximum a posteriori (MAP) probability estimation of $\boldsymbol{\alpha}$ with an identically independent distributed (i.i.d.) Laplacian prior $P(\alpha_i) = 1/2\theta_i \, e^{-|\alpha_i|/\theta_i}$. This Bayesian formulation of sparse coding allows us to gain a better understanding of the duality between the probabilistic and deterministic settings. More specifically, the regularization parameter, which can be set as $\lambda_i = 2\sigma_n^2/\theta_i$, is determined by $\sigma_n^2$, which denotes the noise variance (approximation errors), and $\theta_i$, the standard deviation of the signal (sparse coding coefficients $\alpha_i$) [4]. Modeling the uncertainty on the mean and variance of sparse coefficients has led to two parallel approaches of nonlocal extensions.

### Nonlocal centralized sparse representation

The key idea behind the centralized sparse representation without local locality (NCSR) [5] is to estimate the biased mean of the sparse coding coefficients $\boldsymbol{\alpha}$ by averaging a group of similar nonlocal patches; that is, $\boldsymbol{\beta} = \Sigma_{k \in \Omega} \, \omega_k \boldsymbol{\alpha}_k$, where $\Omega$ denotes the search window to find similar patches, and $\omega_k$ is the linear weight characterizing the similarity between the target and reference patches. Then, the NCSR model is formulated by the following optimization problem:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \| \boldsymbol{x} - D\boldsymbol{\alpha} \|_2^2 + \lambda \sum_i \| \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i \|_p, \tag{2}$$

where the last term contains a nonlocal estimation of the biased mean ($p$ can be one or two) for selected exemplar patches. The uncertainty arising from the estimation of the nonlocal mean can be quantified by the error term $\boldsymbol{e} = \boldsymbol{\alpha} - \boldsymbol{\beta}$. In [5], we have assumed the independence between $\boldsymbol{\beta}$ and $\boldsymbol{e}$ and the i.i.d. Laplace distribution for $\boldsymbol{e}$. It follows that the original nonlinear shrinkage operator for $\boldsymbol{\alpha}$ can be generalized by taking the biased mean into account [4]. Such a generalization allows us to solve (2) by a computationally efficient iterative thresholding algorithm.

### Connecting the Gaussian scale mixture with simultaneous sparse coding

The Gaussian scale mixture (GSM) model has been widely adopted to characterize the spatial variability property of images [3]. The GSM model decomposes the sparse coding coefficients $\boldsymbol{\alpha}$ into the multiplication of a Gaussian vector $\boldsymbol{\beta}$ and a hidden scalar multiplier $\boldsymbol{\theta}$; i.e., $\alpha_i = \theta_i \beta_i$, where the coefficient $\alpha_i$ is Gaussian with a standard derivation of $\theta_i$, and $\theta_i$ is the positive scaling variable (characterizing variance uncertainty). The GSM prior of $\boldsymbol{\alpha}$ can be written as

$$P(\boldsymbol{\alpha}) = \prod_i P(\alpha_i), \; P(\alpha_i) = \int_0^\infty P(\alpha_i \,|\, \theta_i) P(\theta_i) \, d\theta_i . \tag{3}$$

Note that, for most choices of $P(\theta_i)$, it is difficult to compute the MAP estimates of $\alpha_i$. However, we can overcome this difficulty by the joint estimation of $(\alpha_i, \theta_i)$; that is, for a given observation $\boldsymbol{x} = \boldsymbol{D\alpha} + \boldsymbol{n}$, where $\boldsymbol{n} \sim N(0, \sigma_n^2)$, we can formulate the following joint MAP estimation problem:

$$\begin{aligned} (\boldsymbol{\alpha}, \boldsymbol{\theta}) &= \operatorname{argmax} \log P(\boldsymbol{x} \,|\, \boldsymbol{\alpha}, \boldsymbol{\theta}) P(\boldsymbol{\alpha}, \boldsymbol{\theta}) \\ &= \operatorname{argmax} \log P(\boldsymbol{x} \,|\, \boldsymbol{\alpha}) + \log P(\boldsymbol{\alpha} \,|\, \boldsymbol{\theta}) + \log P(\boldsymbol{\theta}), \end{aligned} \tag{4}$$

where $P(\boldsymbol{x} \,|\, \boldsymbol{\alpha})$ is the likelihood term characterized by a Gaussian function with variance $\sigma_n^2$. The prior term $P(\boldsymbol{\alpha} \,|\, \boldsymbol{\theta})$ can be expressed as

$$P(\boldsymbol{\alpha}\,|\,\boldsymbol{\theta}) = \prod_i P(\alpha_i\,|\,\theta_i) = \prod_i \frac{1}{\theta_i\sqrt{2\pi}} \exp\left(-\frac{(\alpha_i-\mu_i)^2}{2\theta_i^2}\right), \quad (5)$$

where $\mu_i$ denotes the biased mean (similar to the weighting coefficient $\beta$ in the NCSR model). With a noninformative prior (as well as Jeffrey's prior) $P(\theta_i) \approx 1/\theta_i$, we can rewrite (5) as

$$(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \underset{\boldsymbol{\alpha},\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_n^2}\| \boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha} \|_2^2 + \sum_i \log(\theta_i\sqrt{2\pi}) + \sum_i \frac{(\alpha_i-\mu_i)^2}{2\theta_i^2} + \sum_i \log\theta_i, \quad (6)$$

where we have used $P(\boldsymbol{\theta}) = \Sigma_i P(\theta_i)$. Noting that Jeffrey's prior is unstable as $\theta_i \to 0$, we replace $\log\theta_i$ with $\log(\theta_i + \epsilon)$, where $\epsilon$ is a small positive number for numerical stability. This equation can then be further translated into the following Bayesian sparse coding problem:

$$(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \underset{\boldsymbol{\alpha},\boldsymbol{\theta}}{\operatorname{argmin}} \| \boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha} \|_2^2 + 4\sigma_n^2 \log(\boldsymbol{\theta} + \epsilon) + \sigma_n^2 \sum_i \frac{(\alpha_i-\mu_i)^2}{\theta_i^2}. \quad (7)$$

A key observation behind simultaneous sparse coding (as well as structured sparsity) [3] is that, for a collection of similar patches, their corresponding sparse coefficients $\boldsymbol{\alpha}$ should be characterized by the same prior—that is, the same $\boldsymbol{\mu}$ (mean) and $\boldsymbol{\theta}$ (variance). Therefore, a matrix extension of (7) can be written as

$$(\boldsymbol{B}, \boldsymbol{\theta}) = \underset{\boldsymbol{B},\boldsymbol{\theta}}{\operatorname{argmin}} \| \boldsymbol{X} - \boldsymbol{D}\Lambda\boldsymbol{B} \|_F^2 + 4\sigma_n^2 \log(\boldsymbol{\theta} + \epsilon) + \sigma_n^2 \| \boldsymbol{B} - \Gamma \|_F^2, \quad (8)$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_m]$ denotes the collection of $m$ similar patches, and $\boldsymbol{A} = \Lambda\boldsymbol{B}$ is the group representation of the GSM model with sparse coefficients. Such a structured sparsity extension of the GSM model allows us to exploit the nonlocal similarity in image signals by low-rank methods [3]. In summary, model-based image reconstruction focuses on the construction of competing image prior models, which is in sharp contrast to the more recently developed data-driven proxy model for the image prior.

### *Bayesian image reconstruction via DUNs*

In the past five years, DCNN-based approaches to image reconstruction have received increasing attention. Assuming a standard degradation mode $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{x}, \boldsymbol{y}$ denotes the original/degraded image pair, and $\boldsymbol{A}$ and $\boldsymbol{n}$ denote the degradation model and additive noise, respectively, we can formulate the following MAP probability estimation problem:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}} \log P(\boldsymbol{x}\,|\,\boldsymbol{y}) = \underset{\boldsymbol{x}}{\operatorname{argmax}} \log P(\boldsymbol{y}\,|\,\boldsymbol{x}) + \log P(\boldsymbol{x}), \quad (9)$$

where $P(\boldsymbol{y}\,|\,\boldsymbol{x})$ and $P(\boldsymbol{x})$ denote the likelihood and prior terms, respectively, which can be written as

$$P(\boldsymbol{y}\,|\,\boldsymbol{x}) \propto \exp\left(-\frac{1}{\sigma_n^2}\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \|_2^2\right), P(\boldsymbol{x}) \propto \exp(-\lambda \boldsymbol{J}(\boldsymbol{x})), \quad (10)$$

where $\sigma_n^2$ is the noise variance, and $\boldsymbol{J}(\boldsymbol{x})$ is the regularization function (e.g., sparsity based [13], nonlocal self-similarity based [7]). It follows that (10) can be rewritten as

$$\boldsymbol{x} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \lambda J(\boldsymbol{x}). \quad (11)$$

An important new insight behind deep learning-based approaches is to unfold the iterative optimization algorithm into a DCNN-based implementation. Such DUNs consist of multiple denoising modules interleaved with back-projection modules to ensure observation consistency. Parallel to model-based approaches, we review the unfolding of two sparse models: denoising prior-driven deep neural networks (DPDNNs) [6] and deep GSM priors [10].

#### DPDNNs

Instead of using an explicitly expressed regularizer, denoising-based image restoration methods [27] use a more complex image prior by decoupling the optimization problem of (11) into one subproblem for the data likelihood term and the other for the prior term. By introducing an auxiliary variable $\boldsymbol{v}$, (11) can be rewritten as

$$(\boldsymbol{x}, \boldsymbol{v}) = \underset{\boldsymbol{x},\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2}\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \lambda J(\boldsymbol{v}), \text{ s.t. } \boldsymbol{x} = \boldsymbol{v}. \quad (12)$$

This equally constrained optimization problem can be converted into the following unconstrained optimization problem:

$$(\boldsymbol{x}, \boldsymbol{v}) = \underset{\boldsymbol{x},\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2}\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \eta \| \boldsymbol{x} - \boldsymbol{v} \|_2^2 + \lambda J(\boldsymbol{v}), \quad (13)$$

which can be solved by alternatively solving two subproblems:

$$\begin{aligned} \boldsymbol{x}^{(t+1)} &= \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \eta \| \boldsymbol{x} - \boldsymbol{v}^{(t)} \|_2^2, \\ \boldsymbol{v}^{(t+1)} &= \underset{\boldsymbol{v}}{\operatorname{argmin}} \eta \| \boldsymbol{x}^{(t+1)} - \boldsymbol{v} \|_2^2 + \lambda J(\boldsymbol{v}). \end{aligned} \quad (14)$$

The $\boldsymbol{x}$ subproblem is quadratic; therefore, it can be solved in closed form by $\boldsymbol{x}^{(t+1)} = \boldsymbol{W}^{-1}\boldsymbol{b}$, where $\boldsymbol{W}$ is a matrix related to the degradation matrix $\boldsymbol{A}$. When the matrix $\boldsymbol{W}$ is large, it is generally impossible to compute its inverse. Instead, the classical conjugate gradient iterative algorithm can be used to calculate $\boldsymbol{x}^{(t+1)}$ but requires many iterations to compute $\boldsymbol{x}^{(t+1)}$. Furthermore, it is difficult to optimize the hyperparameters associated with model-based iterative reconstruction algorithms.

The DUN [9] has offered a promising new direction for both computationally efficient implementation and end-to-end

optimization of algorithm parameters. On the one hand, one can compute $\boldsymbol{x}^{(t+1)}$ with a single step of gradient descent as an approximation to the $\boldsymbol{x}$ subproblem,

$$\begin{aligned}\boldsymbol{x}^{(t+1)} &= x^t - \delta[A^\top(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{y}) + \eta(\boldsymbol{x}^{(t)} - \boldsymbol{v}^{(t)})] \\ &= \bar{\boldsymbol{A}}\boldsymbol{x}^{(t)} + \delta\boldsymbol{A}^\top\boldsymbol{y} + \delta\eta\boldsymbol{v}^{(t)},\end{aligned} \tag{15}$$

where $\bar{\boldsymbol{A}} = [(1-\delta\eta)\boldsymbol{I} - \delta\boldsymbol{A}^\top\boldsymbol{A}]$, and $\delta$ is the parameter that controls the step size. By precomputing $\bar{\boldsymbol{A}}$, the update of $\boldsymbol{x}^{(t)}$ can be computed efficiently. By mimicking the iterative process, updating $\boldsymbol{x}^{(t+1)}$ once is sufficient for $\boldsymbol{x}^{(t)}$ to converge to a local optimal solution. On the other hand, the $\boldsymbol{v}$-subproblem is a proximity operator of $J(\boldsymbol{v})$ calculated at a point $\boldsymbol{x}^{(t+1)}$, whose solution is given by a denoiser; that is, $\boldsymbol{v}^{(t+1)} = f(\boldsymbol{x}^{(t+1)})$. Note that various denoising algorithms can be used, including those that cannot be explicitly expressed by the MAP estimator with $J(\boldsymbol{x})$. Through end-to-end training, both the DCNN-based denoiser $f(\cdot)$ and other network parameters can be jointly optimized, as shown in Figure 2. It is easy to see how the three additive terms on the right side of (15) are assigned to the three separate channels connected by the addition operator (marked with orange) in Figure 2.

## Deep GSM prior

Parallel to structured sparsity or simultaneous sparse coding, we have also extended the work on DUNs to the unfolding of the GSM model in [10], where the variance field of the GSM model is estimated along with the unknown image. Specifically, in [10], we formulate the hyperspectral image (HSI) reconstruction as an MAP estimation problem. With the observed measurements $\boldsymbol{y}$, the target HSI $\boldsymbol{x}$ can be estimated by maximizing the posterior probability, that is,

$$\log p(\boldsymbol{x}|\boldsymbol{y}) \propto \log p(\boldsymbol{y}|\boldsymbol{x}) + \log p(\boldsymbol{x}), \tag{16}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$, and $p(\boldsymbol{x})$ denote the likelihood and the prior distribution, respectively. For the likelihood term, we use a Gaussian function as follows:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2}{2\sigma^2}\right). \tag{17}$$

For the prior term, we propose characterizing each pixel $x_i$ of the HSI by a Gaussian distribution with a nonzero mean and standard deviation $\theta_i$. With a scale prior $p(\theta_i)$ and the independence assumption between $\theta_i$ and $x_i$, we can model the image prior using the following GSM model:

$$p(\boldsymbol{x}) = \prod_i p(x_i), \quad p(x_i) = \int_0^\infty p(x_i|\theta_i)p(\theta_i)d\theta_i, \tag{18}$$

where $p(x_i|\theta_i)$ is the Gaussian distribution; i.e., $p(x_i|\theta_i) = (1/\sqrt{2\pi}\,\theta_i)\exp(-((x_i - u_i)^2/2\theta_i^2))$.

Note that the variance field is embedded in the scale prior $p(\theta_i)$. Instead of modeling $p(\theta_i)$ with an exact prior (e.g., Jeffrey's prior $p(\theta_i) = 1/\theta_i$), we have used the following general form in [10]: $p(\theta_i) \propto \exp(-J(\theta_i))$, where $J(\theta_i)$ is an energy function that plays the role of regularization. Since an analytical expression of $p(x_i)$ is often intractable, we resort to jointly estimating $\boldsymbol{x}$ and $\boldsymbol{\theta}$ by replacing $p(\boldsymbol{x})$ with $p(\boldsymbol{x}, \boldsymbol{\theta})$ in the estimation of MAP. That is,

$$\begin{aligned}(\boldsymbol{x}, \boldsymbol{\theta}) &= \operatorname*{argmax}_{\boldsymbol{x},\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{x}) + \log p(\boldsymbol{x}, \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{x},\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{x}) + \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}).\end{aligned} \tag{19}$$

By substituting the likelihood term $p(x_i|\theta_i)$ and the prior term $p(\theta_i)$ into the MAP estimation, we obtain the following joint objective function:

$$(\boldsymbol{x}, \boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{x},\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \sigma^2 \sum_{i=1}^{N} \frac{1}{\theta_i^2}(x_i - u_i)^2 + 2\sigma^2 J(\boldsymbol{\theta}). \tag{20}$$

Similar to the GSM-simultaneous sparse coding derivation [3], this joint optimization problem can be solved by alternating the optimization of $\boldsymbol{x}$ and $\boldsymbol{\theta}$. With a fixed $\boldsymbol{\theta}$, we can update $\boldsymbol{x}$ by solving the $\boldsymbol{x}$ subproblem:

$$\boldsymbol{x} = \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \sum_{i=1}^{N} w_i(x_i - u_i)^2, \tag{21}$$

where $w_i = \sigma^2/\theta_i^2$, and the mean $u_i$ is updated along with $\boldsymbol{x}$. Similar to the NCSR model [5], we can calculate the weighted average of similar patches as the mean estimate $u_i$. Then, the solution to (21) is given by gradient descent as the following:

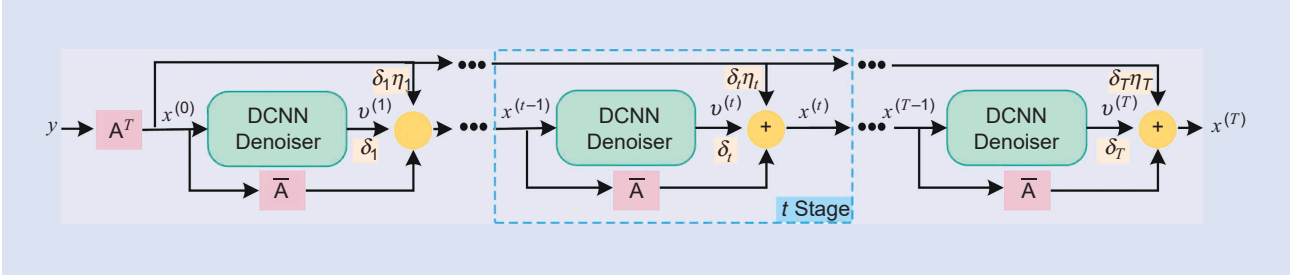



**FIGURE 2.** The unfolding of image reconstruction based on iterative denoising in a DCNN-based implementation [6].

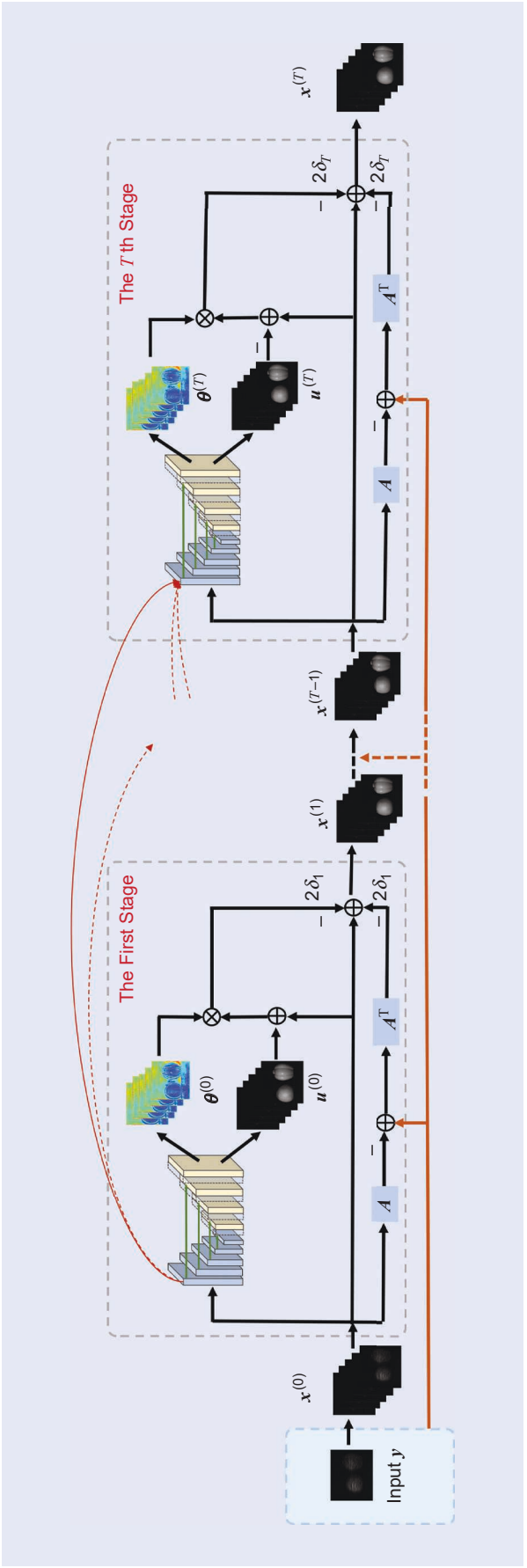



**FIGURE 3.** The general network architecture of the deep GSM prior [10].

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - 2\delta\{\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{y}) + \boldsymbol{w}^{(t)}(\boldsymbol{x}^{(t)} - \boldsymbol{u}^{(t)})\}, \quad (22)$$

where $\boldsymbol{u}^{(t)} = [u_1^t, \ldots, u_N^t]^\top \in \mathbb{R}^N$, $\boldsymbol{w}^{(t)} = [w_1^t, \ldots, w_N^t]^\top \in \mathbb{R}^N$, and $\delta$ is the step size.

For the $\boldsymbol{\theta}$ subproblem with fixed $\boldsymbol{x}$, we translate it into a problem of estimating $\boldsymbol{w}$; that is,

$$\boldsymbol{w} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i=1}^{N} w_i(x_i - u_i)^2 + J(\boldsymbol{w}), \quad (23)$$

where $J(\boldsymbol{w})$ is the regularization term. For some choices, we can derive a closed-form solution; for others, we resort to iterative algorithms. However, the manual design of a proximal operator has its fundamental limitations (e.g., the difficulty of parameter tuning). Conceptually similar to DPDNN [6], we can estimate $\boldsymbol{w}^{(t+1)}$ from $\boldsymbol{x}^{(t+1)}$ using a DCNN as a surrogate prior. For the purpose of the network design, we bridge the $\boldsymbol{x}$ and $\boldsymbol{w}$ subproblems in the following unified framework:

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - 2\delta\{\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{y}) + \mathcal{S}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^{(t)} - \boldsymbol{u}^{(t)})\}, \quad (24)$$

where $\mathcal{S}(\cdot)$ represents the function of the DCNN-based module to estimate $\boldsymbol{w}$—that is, the solution to (23). As shown in Figure 3, we can construct an end-to-end network with $T$ stages corresponding to $T$ iterations to iteratively optimize the unfolded network parameters $\boldsymbol{x}$ and $\boldsymbol{w}$. The three terms on the right side of (24) are mapped to the three channels connected by the addition operator (denoted $\oplus$ with two minus operators above and below) at each stage.

## Present: DUAL for Bayesian image reconstruction

DUAL refers to the emerging class of ideas that attempts to take both model and data uncertainty into account. The uncertainties in the model and the data are also known as *epistemic* and *aleatoric uncertainties*, respectively, in computer vision [12], and they directly affect the generalizability property of DNNs. Recent studies have shown that modeling uncertainty can be tackled in a principled manner under the BDL framework. This line of research can be interpreted as an extension of DUNs because it shows that the classical model-based MAP probability estimation, even after taking the uncertainty of the data into account, can be unfolded into a DCNN-based implementation.

To demonstrate that DUAL offers a unified framework for deep learning-based computational imaging, we present its applications to two imaging tasks in this section: single-image superresolution (SISR) [18] and robust depth completion [30]. The unifying theme is that the construction of the uncertainty estimation module (UEM) leads to the design of a novel loss function that can be incorporated into the training process. Such a unified treatment allows us to pursue the end-to-end optimization of all components, including both UEMs and parameters of DUNs.

### *DUAL for SISR*

There are two classes of uncertainties to consider: aleatoric uncertainty and epistemic uncertainty. The former captures noise

inherent in the observation data, and the latter accounts for the uncertainty of the model about its predictions. Following the standard notation in the literature of SISR, we use $\boldsymbol{y}_i, \boldsymbol{x}_i$ to denote the low-resolution image and the corresponding high-resolution image, respectively. To quantify the aleatoric uncertainty, we use $f(\cdot), \boldsymbol{\theta}_i$ to denote an SISR network and its associated aleatoric uncertainty, leading to the following observation model: $\boldsymbol{x}_i = f(\boldsymbol{y}_i) + \epsilon \boldsymbol{\theta}_i$, where $\epsilon$ observes the Laplace distribution with zero mean and unit variance. It follows from this observation model that

$$p(\boldsymbol{x}_i, \boldsymbol{\theta}_i \mid \boldsymbol{y}_i) = \frac{1}{2\boldsymbol{\theta}_i} \exp\left(-\frac{\| \boldsymbol{x}_i - f(\boldsymbol{y}_i) \|_1}{\boldsymbol{\theta}_i}\right), \tag{25}$$

where $f(\boldsymbol{y}_i)$ and $\boldsymbol{\theta}_i$ denote the SR image (mean) and the uncertainty (variance), respectively. Then, the log-likelihood function can be written as

$$\ln p(\boldsymbol{x}_i, \boldsymbol{\theta}_i \mid \boldsymbol{y}_i) = -\frac{\| \boldsymbol{x}_i - f(\boldsymbol{y}_i) \|_1}{\boldsymbol{\theta}_i} - \ln \boldsymbol{\theta}_i - \ln 2. \tag{26}$$

DUAL aims to estimate not only the SR image (mean) $f(\boldsymbol{y}_i)$ but also the uncertainty (variance) $\boldsymbol{\theta}_i$ simultaneously. As shown in Figure 4, we train two networks with a shared backbone to estimate the log variance $\boldsymbol{s}_i = \ln \boldsymbol{\theta}_i$ together with $f(\boldsymbol{y}_i)$. The maximum likelihood estimation of (26) boils down to the minimization of the following UDL function:

$$\mathcal{L}_{\text{UD}} = \frac{1}{N} \sum_{i=1}^{N} \exp(-\boldsymbol{s}_i) \| \boldsymbol{x}_i - f(\boldsymbol{y}_i) \|_1 + \boldsymbol{s}_i. \tag{27}$$

The loss function $\mathcal{L}_{\text{UD}}$ includes two terms: the first is associated with the fidelity term, and the second prevents the network from predicting infinite uncertainty for all pixels. Those two terms reach equilibrium, but no prior is imposed on the uncertainty estimation.

The limitations of $\mathcal{L}_{\text{UD}}$ have been shown experimentally in [18]. From (27), we can see that the loss function $\mathcal{L}_{\text{UD}}$ has incorporated the variance term $(\boldsymbol{\theta}_i)$ into the divisor of the difference term in absolution. However, a pixel with a large variance (e.g., around edges) will be penalized after the division and will have less impact on the overall loss function. Although this attenuation of pixels with large uncertainty benefits high-level vision tasks, low-level vision tasks, such as SISR, are different. Since pixels with a large uncertainty carry visually important information, they need to be prioritized (instead of attenuated) and given larger (instead of smaller) weights. Such a new insight inspired us to better prioritize pixels with a large uncertainty using a new adaptive weighted loss named *UDL* for SISR.

More specifically, instead of using $\exp(-\boldsymbol{s}_i)$ to attenuate the importance of pixels with a large uncertainty, one can use a monotonically increasing function to prioritize them. Linear scaling is a natural option that leads to the following loss function:

$$\mathcal{L}_{\text{UDL}} = \frac{1}{N} \sum_{i=1}^{N} \hat{s}_i \| \boldsymbol{x}_i - f(\boldsymbol{y}_i) \|_1, \tag{28}$$

where $\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i - \min(\boldsymbol{s}_i)$ is a nonnegative linear scaling function. To prevent the uncertainty value from degenerating into

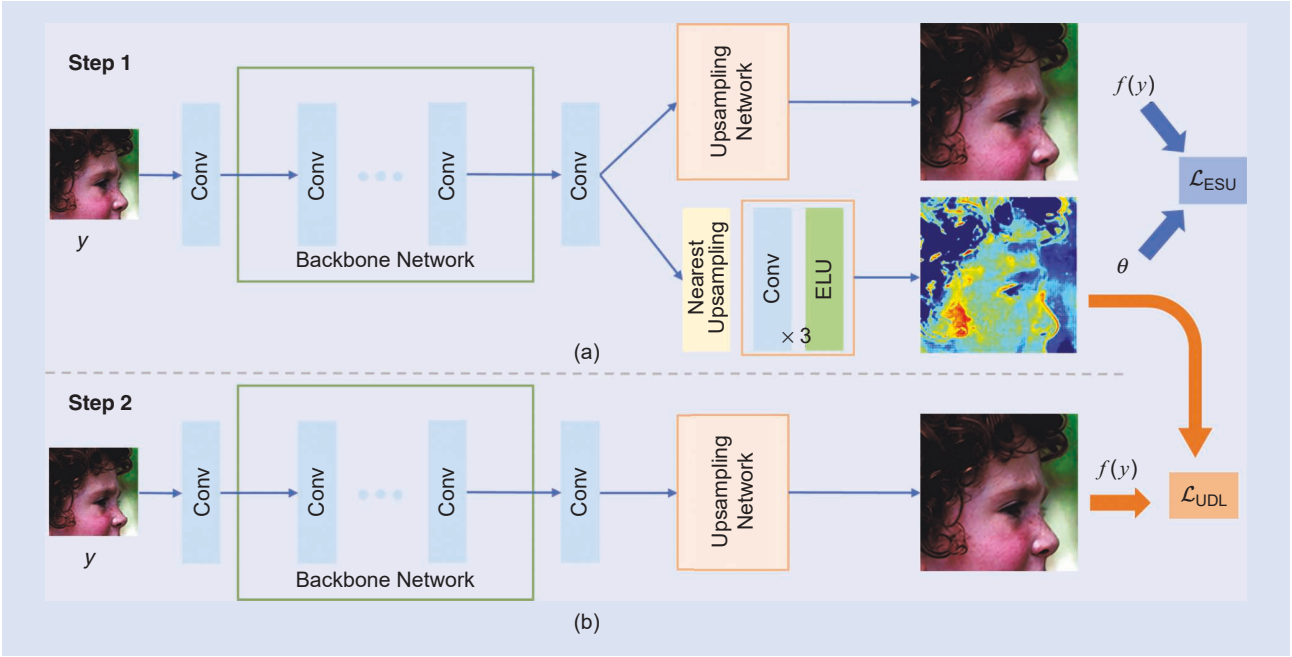



**FIGURE 4.** The training of an SISR network with loss $\mathcal{L}_{\text{UDL}}$ [18]. The training process can be divided into two steps: (a) the first step estimates the uncertainty $\boldsymbol{\theta}$, and (b) the second step generates the final mean value $f(\boldsymbol{y})$. In step 1, shown in (a), the mean value $f(\boldsymbol{y})$ and variance $\boldsymbol{\theta}$ are pretrained by loss $\mathcal{L}_{\text{ESU}}$. During step 2, the mean value network $f(\boldsymbol{y})$ is trained by loss $\mathcal{L}_{\text{UDL}}$, while the inferring variance network $\boldsymbol{\theta}$ is fixed. Conv: convolution; ELU: exponential linear unit.

zeros, the uncertainty estimation result in the first step will be passed to the second step as the attention signal ($\boldsymbol{s} = \ln \boldsymbol{\theta}$). Note that, in $\mathcal{L}_{\text{UDL}}$ loss, the texture and edge pixels with higher uncertainty tend to have larger weights than the pixels in smooth regions, which matches our intuition of prioritizing edges and textures.

To demonstrate the effectiveness of UDL in SISR, we have selected three popular SISR techniques, enhanced deep super-resolution network [15], residual channel attention network [29], and DPDNN [6], as well as the gradient scaling attention model (GRAM) [14] as a baseline for comparison. To our knowledge, GRAM [14] was the only study of data uncertainty in SISR prior to the publication of [18]. It shares a similar observation with uncertainty modeling, but the strategy of the UDL design differs from ours. Table 1 compares the peak signal-to-noise ratio (PSNR)/structural similarity index (SSIM) results for four different network architectures using different loss functions. UDL [18] has consistently achieved a better performance than the original models (without uncertainty) and the baseline (GRAM).

### *DUAL for robust depth completion*

Aleatoric uncertainty that captures the noise inherent in the observations can be further categorized into two classes: homoscedastic and heteroscedastic. Heteroscedastic uncertainty is especially important to the task of depth completion [30] due to the physical limitations of lidar sensors—e.g., lidar often scans the surrounding environment at equally divided angles, resulting in an uneven distribution of depth images. Such an uneven distribution leads to varying densities in different areas, which is the source of heteroscedastic uncertainty. Conventional depth-completion methods average the MSE loss across all pixels, ignoring the issue of heteroscedastic uncertainty. Low-density areas (arising from nonuniform sampling) and outliers often cause the network to overemphasize these areas (i.e., overfitting).

In one of our most recent works [30], we considered a parametric approach to quantifying uncertainty in a depth map by its variance field $\boldsymbol{\Sigma}$. The key idea is to predict an unknown dense depth image $\boldsymbol{X}$ from a sparse depth image $\boldsymbol{Y}$ using a deep learning network $\hat{\boldsymbol{X}} = F(\boldsymbol{Y})$. Then, the problem of depth completion can be formulated by maximizing the posterior probability $P(\boldsymbol{X} \mid \boldsymbol{Y})$. After introducing the uncertainty measure $\boldsymbol{\Sigma}$ ($\sigma$ for a pixel), we can decompose the joint posterior probability into the product of marginals,

$$\begin{aligned} P(\boldsymbol{X}, \boldsymbol{\Sigma} | \boldsymbol{Y}) &= P(\boldsymbol{\Sigma} | \boldsymbol{Y}) P(\boldsymbol{X} | \boldsymbol{\Sigma}, \boldsymbol{Y}) \\ &= \prod p(\sigma_i | y_i) p(x_i | \sigma_i, y_i), \end{aligned} \tag{29}$$

where $x_i$, $\sigma_i$ and $y_i$ denote the pixelwise elements of $\boldsymbol{X}, \boldsymbol{\Sigma}$, and $\boldsymbol{Y}$, respectively. For the likelihood of the uncertainty map $p(\sigma_i \mid y_i)$,, we model it with Jeffrey's prior $P(\sigma_i \mid y_i) \approx (1/\sigma_i)$ based on the intuition of the sparsity on the uncertainty map. For the likelihood term, $p(x_i \mid \sigma_i, y_i)$ can be modeled by a Gaussian distribution observing $\hat{x}_i = F(y_i) \sim N(x_i, \sigma_i)$:

$$p(x_i | \sigma_i, y_i) \approx \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left( -\frac{(\hat{x}_i - x_i)^2}{2\sigma_i^2} \right), \tag{30}$$

where $\hat{x}_i$ denotes a pixel of the image $\hat{\boldsymbol{X}}$. Therefore, we obtain the following MAP estimation problem:

$$\begin{aligned} &\max \sum (\log p(\sigma_i \mid y_i) + \log p(x_i \mid \sigma_i, y_i)) \\ &= \underset{\hat{x}_i, \sigma_i}{\operatorname{argmax}} \sum \left( -2 \log \sigma_i - \frac{(\hat{x}_i - x_i)^2}{2\sigma_i^2} - \frac{1}{2} \log 2\pi \right) \\ &= \underset{\hat{x}_i, \sigma_i}{\operatorname{argmin}} \sum \left( 4 \log \sigma_i + \frac{(\hat{x}_i - x_i)^2}{\sigma_i^2} \right) \\ &= \underset{\hat{x}_i, s_i}{\operatorname{argmin}} \sum (e^{-s_i} (\hat{x}_i - x_i)^2 + 2 s_i), \end{aligned} \tag{31}$$

where $s_i = 2 \log \sigma_i$ ($\sigma_i^2 = e^{s_i}$) models uncertainty about $\hat{x}_i$.

This MAP formulation of uncertainty modeling can be translated into the design of a new UDL function as follows:

$$\mathcal{L}_{\text{UD}} = \frac{1}{N} \sum (e^{-s_i} (\hat{x}_i - x_i)^2 + 2 s_i). \tag{32}$$

From the formula, we observe that the first term will reduce the joint loss of pixels with large differences between the prediction and the ground truth $(\hat{x}_i - x_i)^2$. During the optimization process, the optimizer may increase the uncertainty values

**Table 1. The average PSNR and SSIM results for bicubic downsampling degradation with a scaling factor of ×4 on five benchmark data sets.**

| | | | Set 5 | | Set 14 | | BSD 100 | | Urban 100 | | Manga 109 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Scale** | **Loss** | **PSNR** | **SSIM** | **PSNR** | **SSIM** | **PSNR** | **SSIM** | **PSNR** | **SSIM** | **PSNR** | **SSIM** |
| EDSR-S | ×4 | Original | 31.61 | 0.8862 | 28.22 | 0.7721 | 27.3 | 0.7271 | 25.25 | 0.7575 | 29.31 | 0.8907 |
| | | GRAM | 31.08 | 0.8787 | 27.89 | 0.767 | 27.12 | 0.7229 | 24.81 | 0.7429 | 28.18 | 0.8762 |
| | | $\mathcal{L}_{\text{UDL}}$ | **31.9** | **0.8897** | **28.37** | **0.7755** | **27.4** | **0.7301** | **25.54** | **0.7671** | **29.77** | **0.8967** |
| DPDNN | ×4 | Original | 31.72 | 0.889 | 28.28 | 0.773 | 27.44 | 0.729 | 25.53 | 0.768 | – | – |
| | | GRAM | 31.89 | 0.8913 | 28.37 | 0.7772 | 27.41 | 0.7314 | 25.63 | 0.7708 | 29.70 | 0.9003 |
| | | $\mathcal{L}_{\text{UDL}}$ | **32.2** | **0.8944** | **28.6** | **0.7819** | **27.56** | **0.7356** | **26.09** | **0.7862** | **30.38** | **0.9082** |
| EDSR | ×4 | Original | 32.46 | 0.8968 | 28.8 | 0.7876 | 27.71 | 0.742 | 26.64 | 0.8033 | 31.02 | 0.9148 |
| | | GRAM | 32.32 | 0.8971 | 28.73 | 0.7858 | 27.66 | 0.7395 | 26.35 | 0.7955 | 30.73 | 0.9125 |
| | | $\mathcal{L}_{\text{UDL}}$ | **32.59** | **0.8998** | **28.87** | **0.7889** | **27.78** | **0.7431** | **26.75** | **0.8054** | **31.24** | **0.9167** |
| RCAN | ×4 | Original | 32.54 | 0.8986 | 28.8 | 0.7869 | 27.72 | 0.7418 | 26.6 | 0.8026 | 31.05 | 0.9156 |
| | | $\mathcal{L}_{\text{UDL}}$ | **32.65** | **0.9008** | **28.89** | **0.7896** | **27.81** | **0.7438** | **26.84** | **0.8099** | **31.29** | **0.9198** |

Enhanced deep super-resolution network (EDSR-S) is the EDSR baseline network [15] having 1.5 million parameters. The best performances are shown in bold.

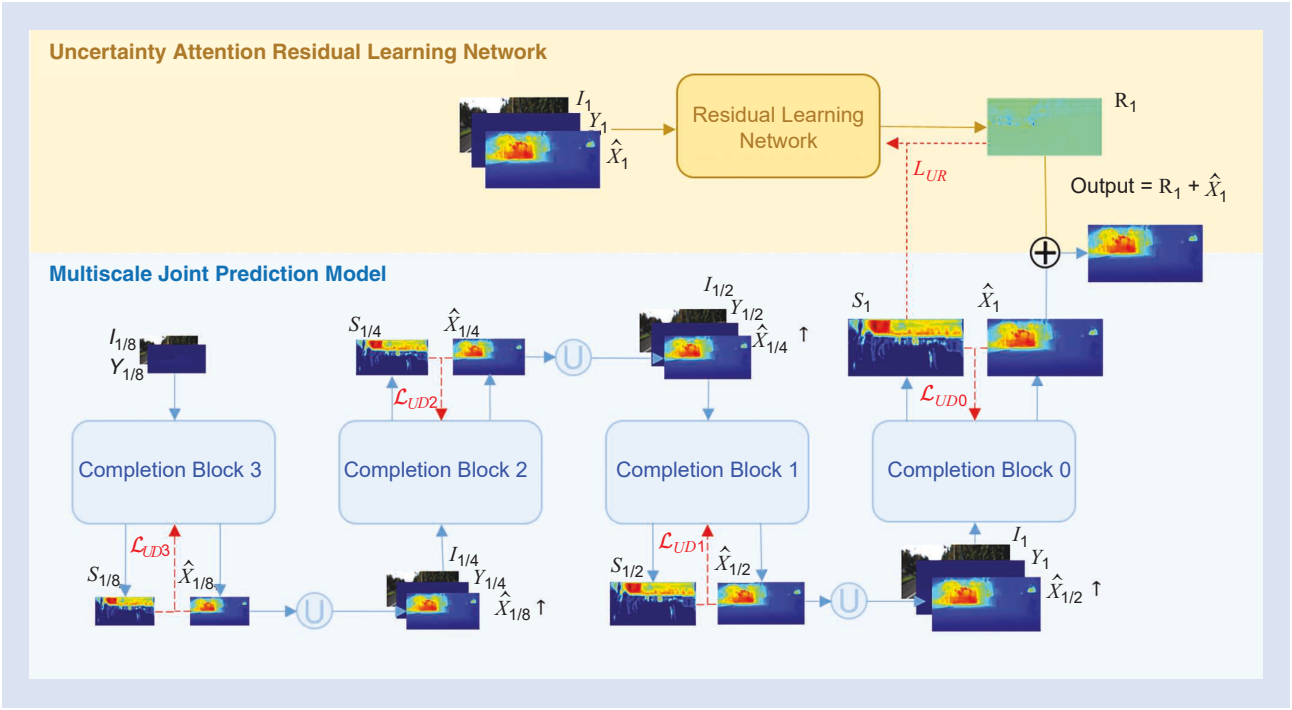



**FIGURE 5.** The robust completion of depth driven by uncertainty [30]. At the bottom (step 1), we jointly predict uncertainty maps and dense depth images using the multiscale joint prediction model. The key idea is to balance the contribution of high-uncertainty regions to the joint loss function. At the top (step 2), an uncertainty attention residual learning network is used to refine the prediction for pixels of high uncertainty. The key idea is to predict the refinement map only for pixels that are uncertain in the first step.

so much that the penalty term $e^{-s_i}$ eventually approaches zero. To balance the first term, the second term limits the growth of uncertainty $s_i$ as a regularization term. As a consequence of balancing, the network will control the contribution of high-uncertainty regions to the joint loss function rather than overfitting these regions.

In the DUAL framework, we can observe that regions with higher depth values often have higher uncertainty values. A new insight brought about by [30] is to use the estimated uncertainty map in the first step to guide the depth-completion refinement procedure in the second step, as shown in Figure 5. In other words, with knowledge about the distribution of high-uncertainty regions, one can tailor the process of optimization for these special regions to achieve an even better completion result. The key idea is to predict the refinement map $R$ for $\hat{X}_1$ only for pixels that are uncertain in the first step. Along this line of reasoning, the loss function associated with uncertainty attention residual learning can be written as

$$\mathcal{L}_{\mathrm{UR}} = \frac{1}{N}\sum s_i \left|(x_i - \hat{x}_i) - r_i\right|,$$
$$\mathcal{L}^2_{\mathrm{UR}} = \frac{1}{N}\sum s_i ((x_i - \hat{x}_i) - r_i)^2, \tag{33}$$

where $r_i$ is the pixel of the predicted residual $R$, and $\hat{x}_i$ is the depth output of the first step. Since optimization of different objective metrics often has conflicting objectives for depth completion, a mixture of forms $L_1$ and $L_2$ is used to build the uncertainty-driven balanced loss function $\mathcal{L}_{\mathrm{URB}}$ as follows:

$$\mathcal{L}_{\mathrm{URB}} = \begin{cases} \mathcal{L}_{\mathrm{UR}}, & N_{\mathrm{epoch}} \text{ is even} \\ \frac{1}{2}(\mathcal{L}_{\mathrm{UR}} + \mathcal{L}^2_{\mathrm{UR}}). & \text{else} \end{cases} \tag{34}$$

As reported in [30], we have verified the effectiveness of UDL functions on the Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago (KITTI) depth-completion benchmark. As shown in Table 2, ours surpasses all other competing methods in mean absolute error (MAE), inverse MAE, and root-mean-square error of the inverse depth

**Table 2. A comparison with other state-of-the-art methods on the KITTI test benchmark.**

| Methods | MAE | iMAE | RMSE | iRMSE |
|---|---|---|---|---|
| NLSPN [19] | 199.59 | 0.84 | 741.68 | 1.99 |
| GuideNet [23] | 218.83 | 0.99 | **736.24** | 2.55 |
| CSPN++ [2] | 209.28 | 0.9 | 743.69 | 2.07 |
| Deep lidar [20] | 226.5 | 1.15 | 758.38 | 2.56 |
| Sparse-to-dense (gd) | 249.95 | 1.21 | 814.73 | 2.8 |
| RGB_guide&certainty | 215.02 | 0.93 | 772.87 | 2.19 |
| UDL [30] (with $L_{URB}$) | 198.09 | 0.85 | 751.59 | **1.98** |
| UDL [30] (with $L_{UR}$) | **190.88** | **0.83** | 795.43 | **1.98** |

Note that uncertainty modeling has led to the best performance in three of the four objective metrics. CSPN: convolutional spatial propagation network; gd: grayscale depth; iMAE: mean absolute error of the inverse depth; NLSPN: non-local spatial propagation networ iRMSE: root-mean-square error of the inverse depth; RGB: red, green, blue.

metrics, demonstrating the superiority of UDL functions. We have also reported some qualitative visual comparison results on the KITTI depth-completion benchmark test data set in [30]. As shown in Figure 6, our results have clear boundaries and recover more details than other depth-completion methods. One salient feature offered by UDL functions is the capability of recovering fine-detailed structures in depth images (e.g., the rearview mirror of the parked car and the vertical pole on the street).

## Future: BDL for image reconstruction

BDL has emerged as a unified framework for tightly integrating deep learning with Bayesian models. In addition to DUN and DUAL, covered in this article, we believe that BDL covers a wider range of ideas, bridging the conventional wisdom of model-based solutions with the new trend of data-driven approaches. From the deep mean shift prior to posterior sampling, there are plenty of room and opportunities to take advantage of theoretically sound ideas originating from Bayesian inference to shed new insight into the new class of deep image priors [24] and plug-and-play priors [27]. Looking ahead, we believe that the following research directions of BDL deserve a systematic study for the next five to ten years: uncertainty-driven kernel estimation for blind image reconstruction (low-level vision), uncertainty-driven transformers for semantic segmentation (middle-level vision), and joint image reconstruction and recognition (high-level vision).

**For low-level vision tasks, modeling real-world degradation has remained a long-standing open problem.**

For low-level vision tasks, modeling real-world degradation has remained a long-standing open problem. The uncertainty factors associated with real-world image degradation are diverse and complex; e.g., the blurring kernel can be motion related or out of focus, spatially invariant, or spatially varying, and the degradation can be associated with adversarial environmental conditions (e.g., atmospheric turbulence or low illumination) or imaging devices (e.g., sensor noise or limited spatial resolution). Despite the progress made for isolated scenarios, there is still a unified framework for systematically taking into account various unknown factors.

Several outstanding open problems remain—e.g., how to properly address the issue of errors in the kernel estimation and noise contamination for blind deconvolution, how to handle spatially varying blur or multiple degradations, and how to unify existing research on blind image denoising/deblurring with blind image superresolution. Some promising results have been reported for the blind reconstruction of face images; much remains to be explored for other image modalities.

In addition to uncertainty modeling, self-supervised learning (SSL) [17] has re-emerged as a compelling framework for representation learning. Several pioneering studies have shown promising results in combining SSL with BDL in low-level vision tasks, such as compressive sensing and medical image reconstruction.

For middle-level vision tasks, such as semantic segmentation, transformer-based approaches (e.g., the shifted window transformer) have shown great potential recently. An uncertainty-guided transformer was recently developed for camouflaged object detection and salient object detection. The motivation behind uncertainty-guided transformer reasoning (UGTR) [25] is to combine a vision transformer with a probabilistic representational model to explicitly reason under uncertainties. The key idea is to first learn a conditional distribution over the transformer output to obtain initial estimates along with associated uncertainties and then reason over these uncertain regions with an attention mechanism to generate final predictions.

Despite the conceptual appeal, the success of UGTR has been limited to the task of detecting objects so far. How can we combine UDL with transformer models for more general vision tasks, such as semantic segmentation? Can we extend the framework of UGTR into semantic segmentation from multimodal data, such as color and depth? How can we take advantage of the vision transformer for efficient uncertainty estimation for semantic segmentation in video?

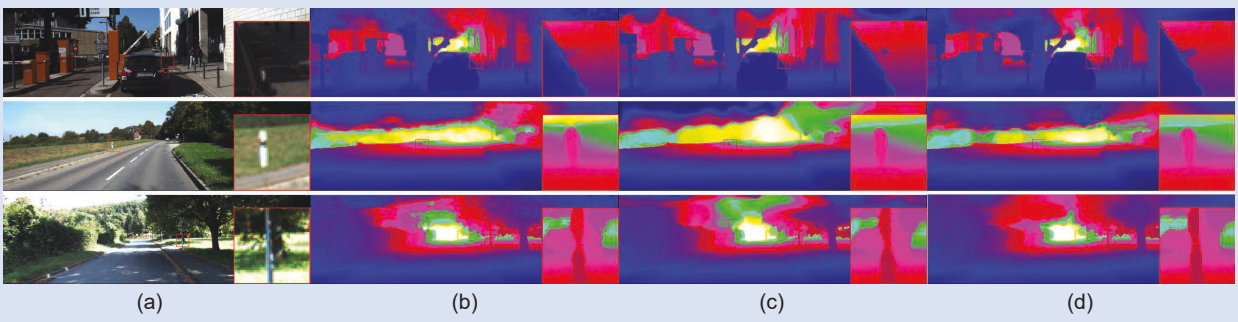



**FIGURE 6.** A comparison of visual quality on the KITTI test benchmark: (a) red, green, blue; (b) CSPN++ [2]; NLSPN [19]; and our UDL [30]. Note that, in the first row, our UDL method [30] is the only one capable of restoring the rearview mirror hidden in the dark background; in the second and third rows, the vertical poles are better recovered by our UDL method [30].

The holy grail of computer vision is teaching a computer to see like humans. One remarkable capability of human vision systems (HVSs) is their robustness and adaptation in challenging adversary environments (e.g., with occlusion and illumination variations). Deep uncertainty learning has shown great potential to improve the robustness of image recognition by feature distillation. However, the problem of robust object recognition has remained largely unsolved. From generalization properties to computational efficiency, there still exists a significant gap between the best invention by humans and innovative discovery by nature (i.e., the evolution and development of HVSs). To fill in this gap, we still need new inspiration from different disciplines.

How can we solve the problem of image reconstruction and object recognition in a closed loop under the framework of BDL? Can we combine BDL and SSL into a unified Bayesian SSL paradigm so that UDL and contrastive loss can be connected? What is the first-order approximation of a truly biologically plausible computer vision system inspired by the organizational principles underlying human visual perception? These important challenges are likely to stimulate further research and attract more young minds to work in this exciting and emerging field.

**Can we extend the framework of UGTR into semantic segmentation from multimodal data, such as color and depth?**

## Conclusion

In this article, we have reviewed the history of image reconstruction from both model-based and learning-based perspectives. From sparse coding to deep learning, Bayesian image reconstruction has evolved into a hybrid framework under which optimization-based solution algorithms for assumed degradation models lead to the principled design of UDL functions in deep learning. In addition to interpretability and transparency, such a marriage between model-based and learning-based paradigms alleviates the burden of handcrafted algorithm parameters by end-to-end optimization. If DUNs mark the bridge connecting traditional optimization-based solution algorithms with fashionable DCNN-based implementations, DUAL is likely to work as a catalyst for uncertainty modeling in unfolded network architectures. Future research on BDL for image reconstruction will continue to benefit from the fruitful interaction between unfolded network architectures and UDL functions.

## Acknowledgments

This work was supported in part by the National Key R&D Program of China under grant 2018AAA0101400 and the Natural Science Foundation of China under grants 61991451, 61632019, 61621005, and 61836008. Xin Li's work is partially supported by the National Science Foundation under grants OAC-1839909 and IIS-1951504.

## Authors

***Weisheng Dong*** (wsdong@mail.xidian.edu.cn) received his B.S. degree in electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004 and his Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2010. He was a visiting student at Microsoft Research Asia, Beijing, China, in 2006. From 2009 to 2010, he was a research assistant with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. In 2010, he joined the School of Artificial Intelligence, Xidian University, Xi'an, 710071, China, as a lecturer, and he has been a professor there since 2016. He was a recipient of the Best Paper Award at the 2010 SPIE Visual Communication and Image Processing. He is currently serving as an associate editor of *IEEE Transactions on Image Processing*. His research interests include inverse problems in image processing, sparse signal representation, and image compression. He is a Member of IEEE.

***Jinjian Wu*** (jinjian.wu@mail.xidian.edu.cn) received his B.Sc. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From September 2011 to March 2013, he was a research assistant at Nanyang Technological University, Singapore. From August 2013 to August 2014, he was a postdoctoral research fellow at Nanyang Technological University. From July 2013 to June 2015, he was a lecturer at Xidian University. Since July 2015, he has been an associate professor with the School of Artificial Intelligence, Xidian University, Xi'an, 710071, China. He served as the special section chair for IEEE Visual Communications and Image Processing 2017 and section chair/organizer/technical program committee member for IEEE International Conference on Multimedia and Expo 2014–2015, Advances in Multimedia Information Processing 2015–2016, IEEE International Conference on Image Processing 2015, and the International Conference on Quality of Multimedia Experience 2016. He was awarded the best student paper of IEEE International Symposium on Circuits and Systems 2013. His research interests include visual perceptual modeling, saliency estimation, quality evaluation, and just-noticeable-difference estimation. He is a Member of IEEE.

***Leida Li*** (ldli@xidian.edu.cn) received his B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. From 2014 to 2015, he was a visiting research fellow with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a senior research fellow from 2016 to 2017. He is currently a professor with the School of Artificial Intelligence, Xidian University, Xi'an, 710071, China. He was a senior program committee member for International Joint Conference on Artificial Intelligence 2019–2020; session chair for ACM International Conference on Multimedia Retrieval 2019 and PCM 2015; and TPC for the American Association for Artificial Intelligence 2019 conference, Association for Computing Machinery International Conference on Multimedia (ACM-MM) 2019–2020, ACM

MM-Asia 2019, the International Conference on Affective Computing and Intelligent Interaction 2019, and the Pacific-Rim Conference on Multimedia 2016. He is an associate editor for the *Journal of Visual Communication and Image Representation* and the European Association for Signal Processing *Journal on Image and Video Processing*. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He is a Member of IEEE.

***Guangming Shi*** (gmshi@mail.xidian.edu.cn) received his B.S. degree in automatic control in 1985, his M.S. degree in computer control in 1988, and his Ph.D. degree in electronic information technology in 2002, all from Xidian University. Since 2003, he has been a professor with the School of Electronic Engineering at Xidian University and, since 2004, the head of the National Instruction Base of Electricians and Electronics. Presently, he is the deputy director of the School of Artificial Intelligence, Xidian University, Xi'an, 710071, China, and the academic leader in the subject of circuits and systems. His research interests include compressed sensing, the theory and design of multirate filter banks, image denoising, low-bit-rate image/video coding, and the implementation of algorithms for intelligent signal processing (using digital signal processing and field-programmable gate arrays). He has authored or coauthored more than 60 research papers. He is a Fellow of IEEE.

***Xin Li*** (xin.li@ieee.org) received his B.S. degree (highest Hons.) in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 1996 and his Ph.D. degree in electrical engineering from Princeton University, Princeton, New Jersey, USA, in 2000. He was a member of technical staff with Sharp Laboratories of America, Camas, Washington, USA, from August 2000 to December 2002. Since January 2003, he has been a faculty member with Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, West Virginia, 26506-6109, USA. He was elected a Fellow of IEEE in 2017.

## References


[1] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numerica*, vol. 28, pp. 1–174, 2019, doi: 10.1017/S0962492919000059.

[2] X. Cheng, P. Wang, C. Guan, and R. Yang, "CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10,615–10,622, doi: 10.1609/aaai.v34i07.6635.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2015, doi: 10.1109/TPAMI.2015.2439281.

[4] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *Proc. CVPR*, 2011, pp. 457–464.

[5] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: A low-rank approach," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 700–711, 2012, doi: 10.1109/TIP.2012.2221729.

[6] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Trans. PAMI*, vol. 41, no. 10, pp. 2305–2318, 2019, doi: 10.1109/TPAMI.2018.2873610.

[7] W. Dong, L. Zhang, R. Lukac, and G. Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1382–1394, 2013, doi: 10.1109/TIP.2012.2231086.

[8] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, 2013, doi: 10.1109/TIP.2012.2235847.

[9] J. R. Hershey, J. L. Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," 2014, *arXiv:1409.2574*.

[10] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep Gaussian scale mixture prior for spectral compressive imaging," in Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2021, pp. 16,216–16,225.

[11] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *Med. Phys.*, vol. 44, no. 10, pp. e360–e375, 2017, doi: 10.1002/mp.12344.

[12] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" 2017, *arXiv preprint:1703.04977*.

[13] K. I. Kim and K. Younghee, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, 2010, doi: 10.1109/TPAMI.2010.25.

[14] C. Lee and K-S. Chung, "Gram: Gradient rescaling attention model for data uncertainty estimation in single image super resolution," in *Proc. 18th IEEE Int. Conf. Mach Learn Appl. (ICMLA)*, 2019, pp. 8–13.

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops (CVPRW)*, 2017, pp. 1132–1140, doi: 10.1109/CVPRW.2017.151.

[16] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Proc. Adv. Neural Inform. Process. Syst.*, 2018, vol. 2018, pp. 1673–1682.

[17] X. Liu *et al.*, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, early access, 2021, doi: 10.1109/TKDE.2021.3090866.

[18] Q. Ning, W. Dong, L. Xin, W. Jinjian, and G. Shi, "Uncertainty-driven loss for single image super-resolution," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021. [Online]. Available: https://papers.nips.cc/paper/2021/hash/88a199611ac2b85bd3f76e8ee7e55650-Abstract.html

[19] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, "Non-local spatial propagation network for depth completion," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 120–136.

[20] J. Qiu *et al.*, "DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 3313–3322, doi: 10.1109/CVPR.2019.00343.

[21] S. Ravishankar, J. Chul Ye, and J. A. Fessler, "Image reconstruction: From sparsity to data-adaptive methods and machine learning," *Proc. IEEE*, vol. 108, no. 1, pp. 86–109, 2019, doi: 10.1109/JPROC.2019.2936204.

[22] J. Sun *et al.*, "Deep ADMM-NET for compressive sensing MRI," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 10–18.

[23] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Trans. Image Process.*, vol. 30, pp. 1116–1129, Aug. 2019, doi: 10.1109/TIP.2020.3040528.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 9446–9454.

[25] F. Yang *et al.*, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4146–4155, doi: 10.1109/ICCV48922.2021.00411.

[26] J. Zhang and B. Ghanem, "ISTA-NET: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 1828–1837, doi: 10.1109/CVPR.2018.00196.

[27] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2021, doi: 10.1109/TPAMI.2021.3088914.

[28] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017, doi: 10.1109/TIP.2017.2662206.

[29] Y. Zhang, L. Kunpeng, L. Kai, W. Lichen, Z. Bineng, and F. Yun, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[30] Y. Zhu, W. Dong, L. Li, J. Wu, X. Li, and G. Shi, "Robust depth completion with uncertainty-driven loss functions," in *Proc. 36th AAAI Artif. Intell. Conf. (AAAI 2022)*.


SP