Emerging Standards Awards from the British Standard Institute.

Thomas Richter (richter@tik.uni-stuttgart.de) received his M.S. degree in physics, his M.S. degree in mathematics, and his Ph.D. degree in mathematical physics in 1995, 1999, and 2000, respectively, from the Technical University of Berlin, Germany. He is a researcher at the TIK Computing Center of the University of Stuttgart, Germany. He has been a member of SC29WG1 since 2003.

Touradj Ebrahimi (Touradj.Ebrahimi@epfl.ch) received his M.S. degree in electrical engineering and his Ph.D. degree in image and video coding from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 1989 and 1992, respectively. He is a professor at EPFL, heading its Multimedia Signal Processing group. He is also the convenor (chair) of the JPEG Standardization Committee.

Rafał K. Mantiuk (rkm38@cam.ac.uk) received his M.S. degree in computer science from the Technical University of Szczecin, Poland, in 2003 and his Ph.D. degree in computer science from the Max Planck Institute for Computer Science, Germany, in 2006. He is a senior lecturer at the Computer Laboratory, University of Cambridge, United Kingdom. He is the author of a popular high dynamic range (HDR) image quality metric, HDR-visual difference predictor-2, and the coauthor of pfstools, software for high-dynamic range image processing.

References

[1] A. Artusi, F. Banterle, T. O. Aydin, D. Panozzo, and O. Sorkine-Hournung, *Image Content Retargeting: Maintaining Color, Tone, and Spatial Consistency.* Boca Raton, FL: CRC, 2016.

[2] A. Artusi, R. Mantiuk, T. Richter, P. Korshunov, P. Hanhart, T. Ebrahimi, and M. Agostinelli, "JPEG XT: A compression standard for HDR and WCG images," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 118–124, 2016.

[3] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. 24th Annu. Conf. Computer Graphics and Interactive Techniques*, 1997, pp. 369–378.

[4] S. Pradeep and C. Aguerrebere, "Practical high dynamic range imaging of everyday scenes: Photographing the world as we see it with our own eyes," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 36–44, 2016.

[5] A. Artusi, R. Mantiuk, R. Thomas, H. Philippe, K. Pavel, A. Massimiliano, T. Arkady, and E. Touradj, "Overview and evaluation of the JPEG XT HDR image compression standard," *Real Time Image Process. J.*, pp. 1–16, Dec. 2015.

[6] R. Thomas, A. Artusi, and E. Touradj, "JPEG XT: A new family of JPEG backward-compatible standards," *IEEE Multimedia Mag.*, vol. 23, no. 3, pp. 80–88, 2016.

[7] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel, "Backward compatible high dynamic range MPEG video compression," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 713–723, 2006.

[8] T. Pouli, A. Artusi, F. Banterle, A. O. Akyüz, H.-P. Seidel, and E. Reinhard, "Color correction for tone reproduction," in *Proc. Color and Imaging Conf. Soc. Imaging Science and Technology*, Nov. 2013, Albuquerque, NM, pp. 215–220.

[9] E. Šikudova, T. Pouli, A. Artusi, A. Ahmet Oğuz, B. Francesco, E. Reinhard, and Z. M. Mazlumoglu, "A gamut mapping framework for color-accurate reproduction of HDR images," *IEEE Trans. Comput. Graph. Appl.*, vol. 36, no. 4, pp. 78–90, 2015.

[10] R. Mantiuk, R. Mantiuk, A. Tomaszewska, and W. Heidrich, "Color correction for tone mapping," *Comput. Graph. Forum*, vol. 28, no. 2, pp. 193–202, 2009.

[11] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. L. Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high dynamic range and standard images," *J. Electron. Imaging*, vol. 24, no. 1, pp. 1–10, 2015.

[12] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 69:1–69:10, Aug. 2008.

Lei Zhang and Wangmeng Zuo

Image Restoration: From Sparse and Low-Rank Priors to Deep Priors

The use of digital imaging devices, ranging from professional digital cinema cameras to consumer grade smartphone cameras, has become ubiquitous. The acquired image is a degraded observation of the unknown latent image, while the degradation comes from various factors such as noise corruption, camera shake, object motion, resolution limit, hazing, rain streaks, or a combination of them. Image restoration (IR), as a fundamental problem in image processing and low-level vision, aims to reconstruct the latent high-quality image from its degraded observation. Image degradation is, in general, irreversible, and IR is a typical ill-posed inverse problem. Due to the large space of natural image contents, prior information on image structures is crucial to regularize the solution space and produce a good estimation of the latent image. Image prior modeling and learning then are key issues in IR research. This lecture note describes the development of image prior modeling and learning techniques, including sparse representation models, low-rank models, and deep learning models.

Relevance

IR plays an important role in many applications, such as digital photography, medical image analysis, remote sensing, surveillance, and digital entertainment. We give an introduction to the major IR techniques developed in past

Digital Object Identifier 10.1109/MSP.2017.2717489
Date of publication: 6 September 2017

1053-5888/17©2017IEEE

years and discuss the future developments. This lecture note can be used as a tutorial to IR methods for senior undergraduate students, graduate students, and researchers in the related areas. The slides associated with this lecture note are available at http://www.comp.polyu.edu.hk/~cslzhang/IR_lecture.pdf.

Prerequisites

Knowledge of statistical signal processing, linear algebra, and convolutional neural networks (CNNs) will be helpful in understanding the content of this lecture note.

Problem statement

Denote by $\boldsymbol{x}$ the latent image and $\boldsymbol{y}$ the degraded observation of it. A typical image degradation model can be written as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \upsilon, \qquad (1)$$

where $\boldsymbol{H}$ denotes the degradation matrix and υ denotes the additive noise. In the literature, υ is often assumed to be additive white Gaussian noise (AWGN) with zero mean and standard deviation σ. Based on the forms of $\boldsymbol{H}$, different IR problems can be defined. For example, in image denoising, $\boldsymbol{H}$ is an identity matrix. In image deblurring, we have $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{k} \otimes \boldsymbol{x}$, where $\boldsymbol{k}$ is the blur kernel and $\otimes$ denotes the two-dimensional (2-D) convolution operator. In image inpainting, $\boldsymbol{H}$ is a diagonal 0-1 matrix. In image superresolution, $\boldsymbol{H}$ is the composition of blurring and downsampling operators.

The linear system in (1) is generally ill posed, i.e., we cannot obtain $\boldsymbol{x}$ by directly solving the linear system. State-of-the-art IR methods exploit image prior information and optimize an energy function to estimate the desired image $\boldsymbol{x}$. From the Bayesian perspective, (1) defines a likelihood function $P(\boldsymbol{y}|\boldsymbol{x}) = exp\{-\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2/2\sigma^2\}$ of $\boldsymbol{x}$. Given the image prior $P(\boldsymbol{x})$, we can estimate the unknown latent image $\boldsymbol{x}$ from the observation $\boldsymbol{y}$ by maximizing the posterior probability $P(\boldsymbol{x}|\boldsymbol{y})$, and the widely used maximum a posterior (MAP) model is

$$\begin{aligned}\hat{\boldsymbol{x}} &= \arg\max\nolimits_{\boldsymbol{x}}\{\log P(\boldsymbol{x}|\boldsymbol{y})\\ &\propto (\log P(\boldsymbol{y}|\boldsymbol{x}) + \log P(\boldsymbol{x}))\}\\ &= \arg\min\nolimits_{\boldsymbol{x}}\{1/2\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 + \lambda R(\boldsymbol{x})\},\end{aligned} \qquad (2)$$

where $R(\boldsymbol{x}) = -\log P(\boldsymbol{x})$ denotes the regularization term and λ is the regularization parameter. Under the MAP framework, one key problem is how to model the image priors $P(\boldsymbol{x})$ [or regularizers $R(\boldsymbol{x})$]. Successful prior models include sparse and low-rank priors. Recently, deep learning techniques have also been used to learn discriminative prior models.

Solutions

Sparse representation

Many IR methods exploit the sparsity prior of natural images. For example, image gradient exhibits long-tailed distribution, with which the total variation methods have been widely used for solving IR problems. The wavelet transform of an image has sparsely distributed coefficients, and, therefore, soft-thresholding in wavelet domain is a simple yet effective denoising technique. An image patch can be well represented as the linear combination of a few atoms sparsely selected from a dictionary. The sparse representation (also known as *sparse coding*)-based methods encode an image patch over an overcomplete dictionary $\boldsymbol{D}$ with ℓ_1-norm (or ℓ_p-norm, $0 \leq p \leq 1$) sparsity regularization on the coding vector, i.e., $\min_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1 \;\; s.t. \, \boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$, leading to a general sparse representation-based IR model

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{D}\boldsymbol{\alpha}\| + \lambda\|\boldsymbol{\alpha}\|_1, \quad (3)$$

which turns the estimation of $\boldsymbol{x}$ in (2) into the estimation of $\boldsymbol{\alpha}$. Table 1 summarizes the major steps in sparse representation-based IR.

Figure 1 shows an example of image denoising by sparse representation. One can see that the noise is rapidly removed during the iteration, and the image is well reconstructed in five iterations. The success of sparse representation in IR can be explained from different perspectives. First, from the Bayesian perspective, it solves a MAP problem with a good sparsity prior. Second, it has neuroscience explanations that the receptive fields of simple cells in primary visual cortex can be characterized as spatially localized, oriented, and bandpass. Third, from the perspective of compressed sensing, image patches are K-sparse signals, which can be well reconstructed using sparse optimization.

The selection of a dictionary

The dictionary $\boldsymbol{D}$ plays an important role in sparse representation-based IR. Early methods usually adopt some analytical dictionaries, such as discrete cosine transform bases, wavelets, curvelets, or the concatenation of them. Nonetheless, these analytically designed dictionaries have limited capability in matching the complex natural image structures. More atoms have to be selected to represent the given image, making the representation less sparse. To address this issue, researchers proposed to learn dictionaries from natural images. Given a set of training samples $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]$, where $\boldsymbol{y}_i$ is a vectorized image patch, dictionary learning aims to learn a dictionary $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m]$ from $\boldsymbol{Y}$, where $\boldsymbol{d}_j$ is an atom and $m < n$, such that $\boldsymbol{Y} \approx \boldsymbol{D}\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n]$ is the set of sparse codes. One classical dictionary learning method is the K-SVD [1] algorithm, which imposes ℓ_0-norm sparsity on each coding vector $\boldsymbol{\alpha}_i, i = 1, 2, \ldots, n$.

Table 1. Major steps in sparse representation-based IR.

1) Partition the degraded image into overlapped patches.

2) For each patch, solve the nonlinear ℓ_1-norm sparse coding problem in (3).

3) Reconstruct each patch by $\hat{\boldsymbol{x}} = \boldsymbol{D}\hat{\boldsymbol{\alpha}}$.

4) Put the reconstructed patch back to the original image. For overlapped pixels between patches, average them.

5) Iterate the above procedures for a better restoration.

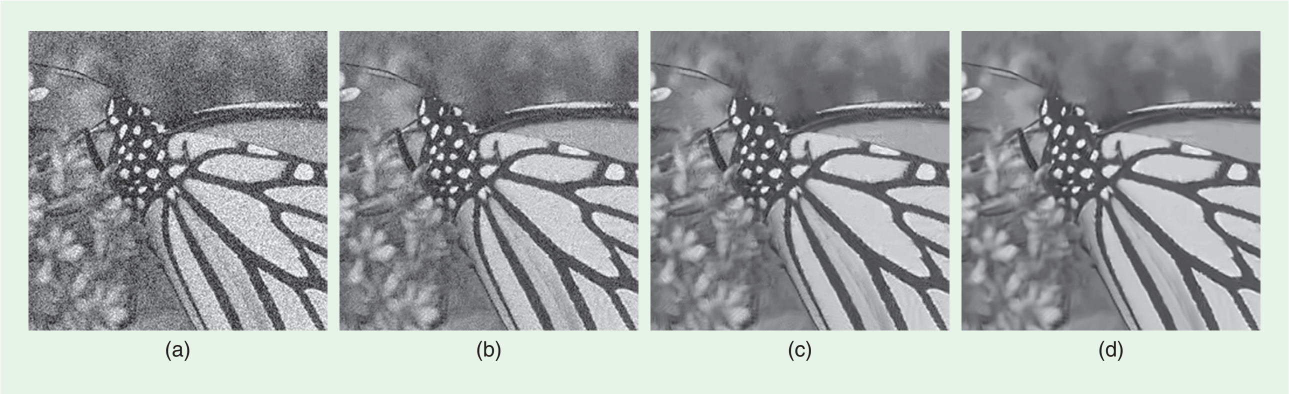

FIGURE 1. The (a) noisy image and the denoised images in iterations (b) one, (c) three, and (d) five.

The success of K-SVD in denoising inspired many other works, such as multiscale dictionary learning, double sparsity, adaptive PCA dictionaries, and semicoupled dictionary learning. Sparse representation models with learned dictionaries often work better than analytically designed dictionaries because the learned dictionaries are more adaptive to specific tasks/data and more flexible to represent the image structures.

Nonlocally centralized sparse representation

Apart from sparsity prior, another widely used image prior is the nonlocal self-similarity (NSS) prior. As shown in Figure 2, natural images usually contain many repetitive local patterns. The similar patches to a given local patch can be spatially far from it. Coupled with sparse representation techniques, NSS-based methods have achieved state-of-the-art performance in many IR problems. One representative work is the so-called nonlocally centralized sparse representation (NCSR) model [4].

The sparse representation of a latent image $\boldsymbol{x}$ over a dictionary $\boldsymbol{D}$ can be equivalently written as

$$\boldsymbol{\alpha}_x = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1, \text{ s.t. } \|\boldsymbol{x} - \boldsymbol{D\alpha}\|_2^2 \le \epsilon, \quad (4)$$

where ϵ is a small positive real number. However, in practice, what we have is the degraded image $\boldsymbol{y}$ instead of the latent image $\boldsymbol{x}$. The sparse code of $\boldsymbol{y}$ is

$$\boldsymbol{\alpha}_y = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1, \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{HD\alpha}\|_2^2 \le \epsilon. \quad (5)$$

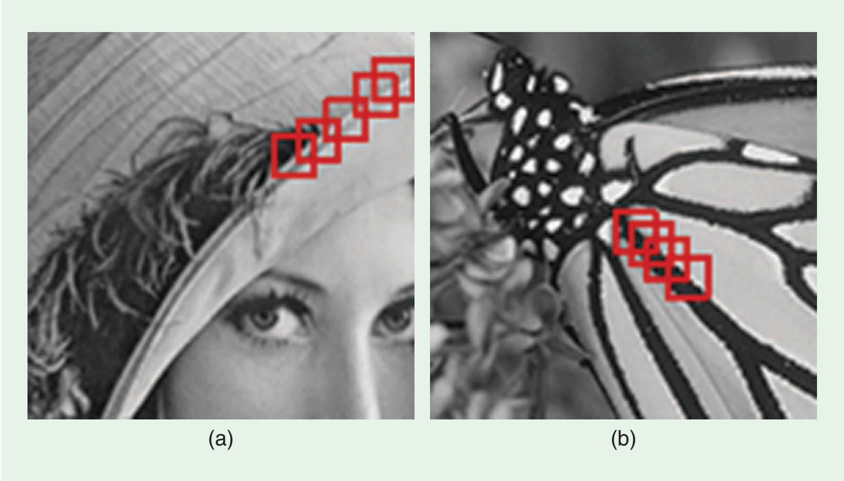

FIGURE 2. The image NSS. The red squares indicate the similar local patterns in an image. (a) The *Lenna* image and (b) the *Butterfly* image.

Clearly, $\boldsymbol{\alpha}_y$ will be different from $\boldsymbol{\alpha}_x$, and we call the difference between them *sparse coding noise* (*SCN*) $\boldsymbol{\upsilon}_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_y - \boldsymbol{\alpha}_x$. To better reconstruct $\boldsymbol{x}$ from $\boldsymbol{y}$, we should reduce the SCN $\boldsymbol{\upsilon}_{\boldsymbol{\alpha}}$ as much as possible. Suppose that we have some good estimation of $\boldsymbol{\alpha}_x$, denoted by $\hat{\boldsymbol{\alpha}}_x$, then one good (unconstrained) sparse coding model to suppress $\boldsymbol{\upsilon}_{\boldsymbol{\alpha}}$ is

$$\boldsymbol{\alpha}_y = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{y} - \boldsymbol{HD\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}_x\|_1\}, \quad (6)$$

which enforces the output code $\boldsymbol{\alpha}_y$ to be close to $\hat{\boldsymbol{\alpha}}_x$.

Now the problem turns to how to obtain $\hat{\boldsymbol{\alpha}}_x$ since $\boldsymbol{x}$ is not available. Without additional information, an unbiased estimate of $\boldsymbol{\alpha}_x$ will be the mathematical expectation of it, i.e., $\hat{\boldsymbol{\alpha}}_x = E[\boldsymbol{\alpha}_x]$. Meanwhile, it is empirically found that the SCN $\boldsymbol{\upsilon}_{\boldsymbol{\alpha}}$ has zero mean and is Laplacian distributed, which leads to $\hat{\boldsymbol{\alpha}}_x = E[\boldsymbol{\alpha}_x] \approx E[\boldsymbol{\alpha}_y]$. Therefore, for each image patch $\boldsymbol{x}_i$, we can estimate its $E[\boldsymbol{\alpha}_x]$ as the nonlocal means of $\boldsymbol{\alpha}_y$: $\boldsymbol{\mu}_i = \Sigma_{j \in C_j} \omega_{i,j} \boldsymbol{\alpha}_{i,j}$, where $\omega_{i,j} = \exp(\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_{i,j}\|_2^2 / h)/W$, $\hat{\boldsymbol{x}}_i$ is the current estimation of $\boldsymbol{x}_i$, $\hat{\boldsymbol{x}}_{i,j}$ are the nonlocal similar patches to $\hat{\boldsymbol{x}}_i$ in a search range C_i, and W is a normalization factor. Finally, the NCSR model becomes

$$\boldsymbol{\alpha}_y = \arg\min_{\boldsymbol{\alpha}} \left\{\|\boldsymbol{y} - \boldsymbol{HD\alpha}\|_2^2 + \lambda \sum_{i=1}^{N} \|\boldsymbol{\alpha}_i - \boldsymbol{\mu}_i\|_1\right\}, \quad (7)$$

which can be easily solved iteratively. The NCSR model naturally integrates

NSS into the sparse representation framework and shows competitive performance in different IR applications, including denoising, debluring, and superresolution [4].

Low-rank minimization

The sparse representation models stretch an image patch to a vector and encode it over a dictionary of one-dimensional (1-D) atoms. With the NSS prior, we can have a group of similar patches as input. Group sparsity models have been proposed to encode a group of correlated patches, whereas they are still a type of 1-D sparse coding model. An alternative way is to format those similar patches as a matrix with each column being a stretched patch vector, and exploit the low-rank prior of this matrix for IR.

The rank of a data matrix $\boldsymbol{X}$ counts the number of nonzero singular values of it, which is NP-hard to minimize. Alternatively, the nuclear norm of $\boldsymbol{X}$, defined as the ℓ_1-norm of its singular values $\|\boldsymbol{X}\|_* = \Sigma_i \|\sigma_i(\boldsymbol{X})\|_1$, is a convex relaxation of matrix rank function. The low-rankness of $\boldsymbol{X}$ can be viewed as a 2-D sparsity prior. It encodes the input 2-D data matrix over a set of rank-1 basis matrices and assumes its singular values to be sparsely distributed, i.e., it has only a few nonzero or significant singular values.

Let $\boldsymbol{Y}$ be a matrix of degraded image patches. The latent low-rank matrix $\boldsymbol{X}$ can be estimated form $\boldsymbol{Y}$ via the following nuclear norm minimization (NNM) problem:

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_*. \quad (8)$$

Cai et al. [2] showed that (8) has a closed-form solution

$$\hat{\boldsymbol{X}} = \boldsymbol{U}\mathcal{S}_{\frac{\lambda}{2}}(\Sigma)\boldsymbol{V}^T, \quad (9)$$

where $\boldsymbol{Y} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ is the SVD of $\boldsymbol{Y}$ and $\mathcal{S}_{\frac{\lambda}{2}}(\Sigma)_{ii} = \max(\Sigma_{ii} - (\lambda/2), 0)$ is the singular value thresholding operator.

Weighted NNM

The NNM mentioned previously has shown interesting results on image and video denoising. As can be seen from (9), however, it shrinks all the singular values equally by the threshold λ, ignoring the different significances of matrix singular values. It is known that the larger singular values can be more important to represent the latent data in many applications. In [6], a weighted nuclear norm is defined

$$\|\boldsymbol{X}\|_{w,*} = \sum_i \|w_i \sigma_i(\boldsymbol{X})\|_1, \quad (10)$$

where w_i is the weight assigned on singular value $\sigma_i(\boldsymbol{X})$. A weighted NNM (WNNM) model is then presented [6] to recover the latent data matrix $\boldsymbol{X}$ from $\boldsymbol{Y}$

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \|\boldsymbol{X}\|_{w,*}. \quad (11)$$

Different from the convex NNM model in (8), the WNNM model in (11) becomes nonconvex. Fortunately—and interestingly—it is proved in [6] that WNNM still has a globally optimal solution.

It is also shown in [6] that, if the weights satisfy $0 \le w_1 \le \cdots \le w_n$, the nonconvex WNNM problem has a closed form optimal solution

$$\hat{\boldsymbol{X}} = \boldsymbol{U}\mathcal{S}_{\frac{w}{2}}(\Sigma)\boldsymbol{V}^T, \quad (12)$$

where $\boldsymbol{Y} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ is the SVD of $\boldsymbol{Y}$ and $\mathcal{S}_{\frac{w}{2}}(\Sigma)_{ii} = \max(\Sigma_{ii} - (w_i/2), 0)$.

The above conclusion is very useful. In many IR applications, the weights can be easily set as nonascending [6], and, therefore, WNNM has a closed-form solution, which makes the minimization process efficient. Figure 3 illustrates the WNNM-based image denoising scheme. For each noisy patch, we search its nonlocal similar patches to form the matrix $\boldsymbol{Y}$. Then we solve the WNNM problem in (11) to estimate the clean patches $\boldsymbol{X}$. Once the estimated clean patch is put back into the image, the noise is reduced. Such procedures are repeated several times to obtain the denoised image. WNNM has shown state-of-the-art denoising results.

Deep prior learning

The sparse representation and low-rank minimization-based IR methods discussed above are model-based optimization schemes, where a model (objective function) is built based on the image degradation process and the available image priors, and the desired image is reconstructed by finding the optimal solution of the model. Such models can be generally written as

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} F(\boldsymbol{x}, \boldsymbol{y}) + \lambda R(\boldsymbol{x}), \quad (13)$$

where $F(\cdot)$ is the data fidelity term (e.g., $F(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{Hx}\|_2^2$) and $R(\cdot)$ is the regularization term (or prior term).

Another category of IR methods is the so-called discriminative learning methods, which learn a compact inference or a mapping function from a training set of degraded-latent image

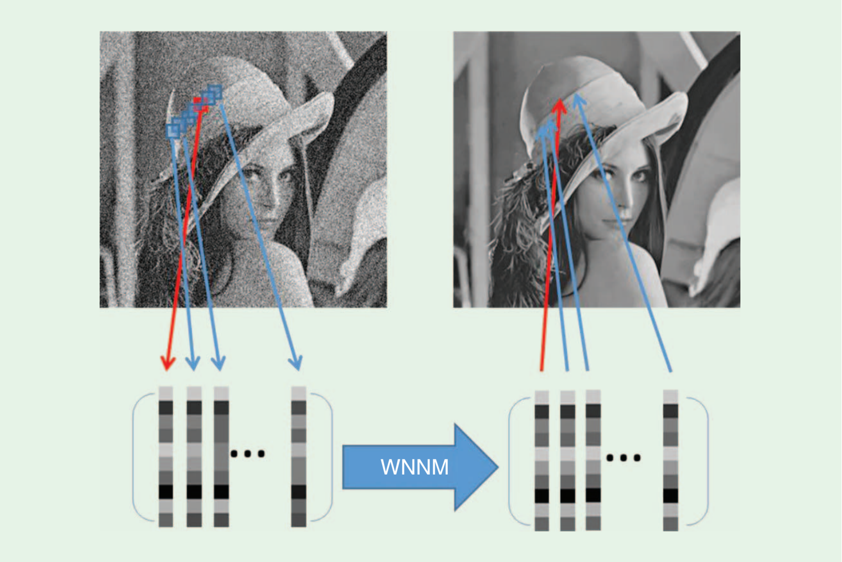

FIGURE 3. An example of WNNM-based image denoising.

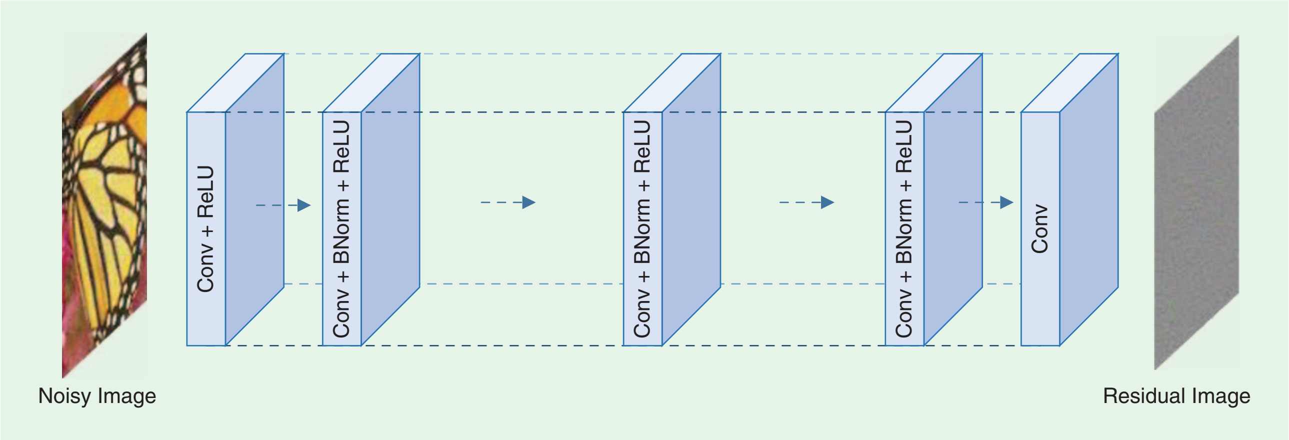

FIGURE 4. A residual learning-based image denoising framework. Conv: convolution, ReLU: rectified linear unit, and BNorm: batch normalization.

pairs. The general model of discriminative learning methods can be written as

$$\min_{\Theta} \text{loss}(\hat{\boldsymbol{x}}, \boldsymbol{x}), \quad \text{s.t. } \hat{\boldsymbol{x}} = \mathcal{F}(\boldsymbol{y}, \boldsymbol{H}; \Theta), \tag{14}$$

where $\mathcal{F}(\cdot)$ is the inference or mapping function with parameter set Θ and $\text{loss}(\cdot)$ is the loss function to measure the similarity between output image $\hat{\boldsymbol{x}}$ and ground-truth image $\boldsymbol{x}$. The recently developed deep learning-based IR methods [3], [7]–[10] are typical discriminative learning methods, where $\mathcal{F}(\cdot)$ is a deep CNN.

Deep CNNs for IR

One of the first CNN-based IR methods is the superresolution CNN (SRCNN) method [3] for single image superresolution (SISR). It is actually modestly deep because it has only two hidden layers. A truly deep CNN for SISR was developed in [7]. The so-called very deep superresolution (VDSR) method first initializes the low-resolution (LR) image to a high-resolution (HR) image (e.g., by bicubic interpolation) and then learns a CNN to predict the residual between the initialized HR image and the ground-truth image. VDSR shows highly competitive peak signal-to-noise ratio (PSNR) results, and it demonstrates that a single CNN can perform SISR with multiple scaling factors.

VDSR enlarges the LR input to the same size of the HR image before it goes through a CNN. This not only restricts the area of receptive fields but also increases the cost of convolution operations in CNN. A more efficient solution is to directly predict the missing HR pixels from the LR image, as proposed in [9]. The so-called efficient subpixel CNN (ESPCNN) learns to predict the feature maps of each subimage, and aggregates the subimages into the final HR image.

Most existing SISR methods use the mean squared error (MSE) as the loss function to optimize the network parameters Θ. Minimizing MSE tends to produce the mean image of possible observations. As a result, the output HR image can be oversmoothed without a good perceptual quality. Inspired by the great success of the recently developed generative adversarial nets (GANs) [5]—a perceptual loss function, which consists of a content loss (i.e., MSE) and a GAN-based adversarial loss—is proposed in [8] to generate the HR images. Though the PSNR index is not very high, the so-called superresolution with GAN method produces perceptually very pleasant SISR results.

CNNs have also been successfully used in image denoising. A residual learning-based image denoising method, called *DnCNN*, is proposed in [10], whose framework is shown in Figure 4. The network is learned to predict the noise (i.e., residual) corrupted in the image, and the denoised image can be obtained by subtracting the predicted noise from the noisy image. It is shown in [10] that batch normalization is very useful for Gaussian noise removal since it enforces the output in each layer to be Gaussian-like distributed. The DnCNN method can also handle more general noise, such as the compression errors and the interpolation errors, using a single network. Figure 5 shows an example. The input image is partly noise corrupted, partly interpolated, and partly compressed. One can see that the trained DnCNN can handle the three tasks simultaneously with very good performance.

Learn a deep denoiser prior for general IR tasks

Though CNN has achieved very competitive performance compared with the model-based optimization method in IR, it has limitations, especially on the capacity of generality. In most of the CNN-based SISR methods, the downsampling kernel is assumed to be the bicubic kernel. When applying the trained CNN to an LR image produced by a different kernel, the result will become much less satisfactory. Similar things will happen for image deblurring problems, where the blur kernels can be very different but it is hardly possible to train a CNN for each instantiation of the blur kernel. Table 2 summarizes the pros and cons of model-based optimization methods and deep CNN-based methods for IR.

Compared with the deep CNN-based methods, model-based optimization methods have better generality. Given

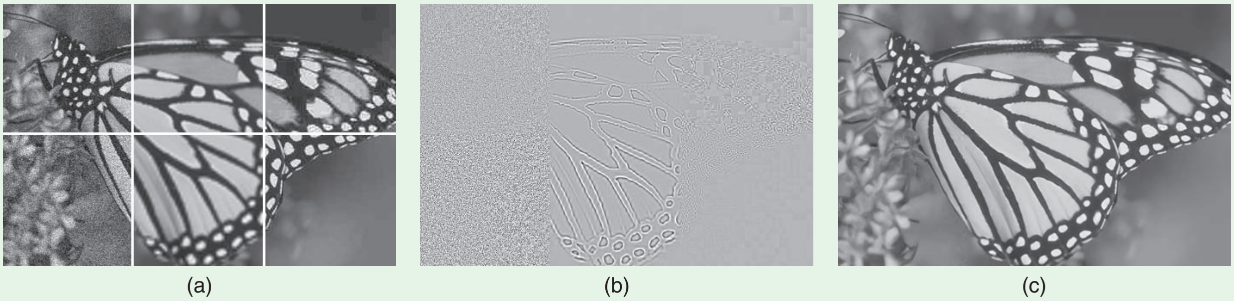

FIGURE 5. An example of the capacity of DnCNN for three different tasks. (a) Input image, (b) output residual image, and (c) restored image. Parts of (a) are corrupted with noise levels 15 (top left) and 25 (bottom left), bicubically interpolated with upscaling factors 2 (top middle) and 3 (bottom middle), and JPEG compressed with quality factors 10 (top right) and 30 (bottom right).

Table 2. Pros and cons of model-based optimization methods and deep CNN-based methods for IR.

	Model-based optimization methods	Deep CNN-based methods
Pros	✓ General to handle different IR problems ✓ Clear physical meanings	✓ Data driven end-to-end learning ✓ Can be very efficient in the testing stage
Cons	× The handcrafted priors may not be strong enough × The optimization process can be time-consuming	× The generality of learned models is limited × The interpretability of learned models is limited

the image prior $R(\boldsymbol{x})$, one can easily extend the optimization algorithm to solve any IR tasks with different degradation matrices $\boldsymbol{H}$. One natural question is whether we can integrate the CNN-based prior learning with model-based optimization to develop a general and powerful IR method.

One obstacle to the integration of CNN and model-based methods lies in that the image prior $R(\boldsymbol{x})$ is not explicitly given in CNN-based methods. Fortunately, in many model-based optimization methods, what we need is not the explicit $R(\boldsymbol{x})$ but a powerful denoiser. One can readily plug a CNN denoiser into the optimization methods for solving various IR tasks. An example of such a scheme, called *IRCNN*, has been developed in [11].

By means of half-quadratic splitting (HQS), we can introduce an auxiliary variable z and reformulate (2) as the following unconstrained optimization problem to alternatingly solve $\boldsymbol{x}$ and z:

$$\mathcal{L}_\mu(\boldsymbol{x},z) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 + \lambda R(z) + \frac{\mu}{2}\|\boldsymbol{x} - z\|_2^2, \qquad (15)$$

where μ is a nondescending penalty parameter with iterations. $\boldsymbol{x}$ and z are updated as

$$\boldsymbol{x}^{k+1} = (\boldsymbol{H}^T\boldsymbol{x} + \mu \boldsymbol{I})^{-1}(\boldsymbol{H}^T\boldsymbol{y} + \mu z^k), \quad (16)$$

$$z^{k+1} = \arg\min_z \frac{1}{2(\sqrt{\lambda/\mu})^2}\|z - \boldsymbol{x}^{k+1}\|_2^2 + R(z). \qquad (17)$$

One can see that the z-subproblem in (17) is a denoising problem, and a CNN denoiser can be trained to solve it

$$z^{k+1} = \text{CNN\_Denoiser}(\boldsymbol{x}^{k+1}, \sqrt{\lambda/\mu}). \qquad (18)$$

The CNN denoiser and HQS-based IRCNN method [11] is illustrated in Figure 6. The noise level σ of the CNN

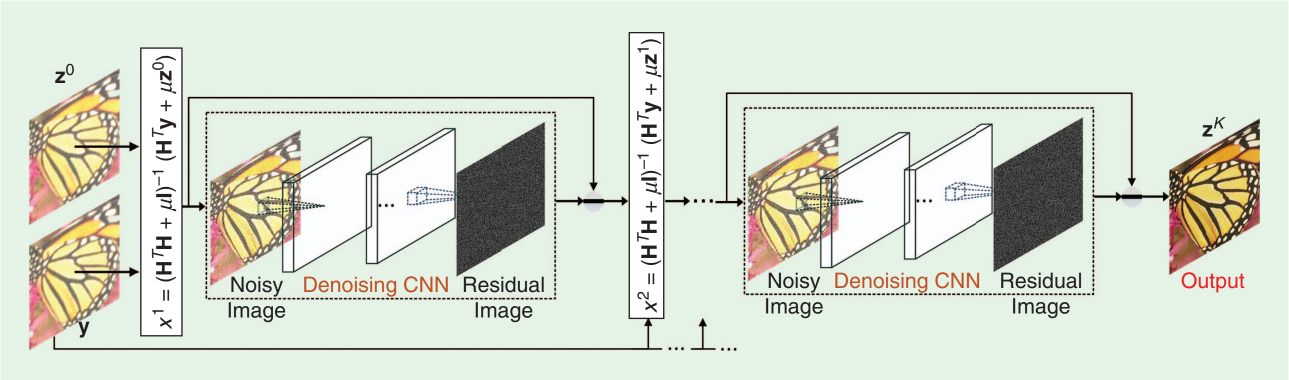
$\mathbf{z}^0$
$\mathbf{y}$
$x^1 = (\mathbf{H}^T\mathbf{H} + \mu\mathbf{I})^{-1}(\mathbf{H}^T\mathbf{y} + \mu\mathbf{z}^0)$
Noisy Image
Denoising CNN
Residual Image
$x^2 = (\mathbf{H}^T\mathbf{H} + \mu\mathbf{I})^{-1}(\mathbf{H}^T\mathbf{y} + \mu\mathbf{z}^1)$
Noisy Image
Denoising CNN
Residual Image
$\mathbf{z}^K$
Output

FIGURE 6. Incorporating a CNN denoiser with HQS for IR.

denoiser is related to the fidelity term and varies along with the iterations. Twenty five denoisers are trained in the noise level range [0, 50] with step size 2. In the test stage, the CNN denoiser whose noise level σ is nearest to $\sqrt{\lambda/\mu}$ is chosen in each iteration. The architecture of the CNN denoiser is illustrated in Figure 7.

Figure 8 shows the deblurring results on image *Leaves*. One can see that IRCNN [11] is very promising in recovering image sharpness and naturalness.

Figure 9 presents the SISR results on the image *Butterfly*. Since VDSR is trained on images downsampled with the bicubic kernel, it cannot be directly extended to other downsampling settings without retraining. In contrast, the

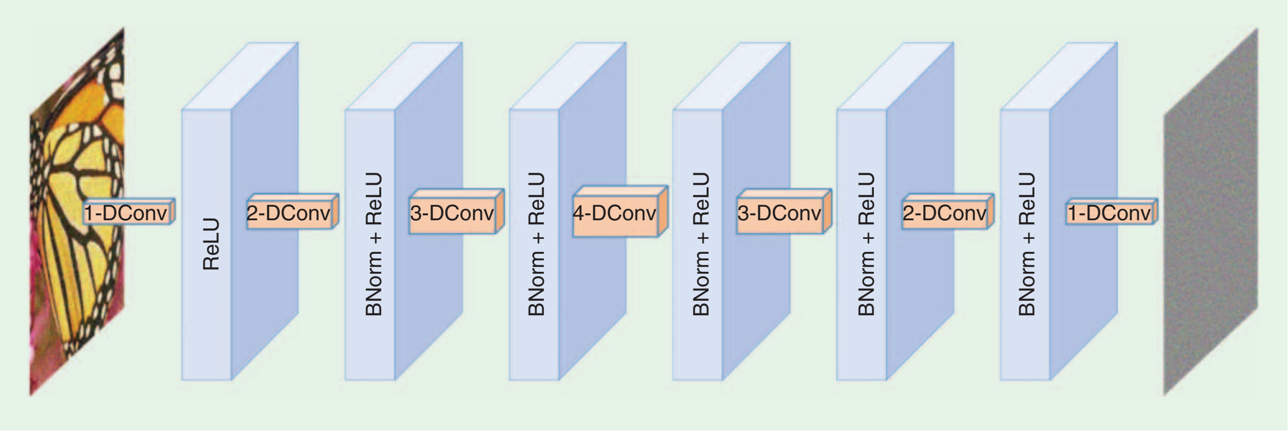

FIGURE 7. The architecture of the CNN denoiser. *s*-DConv: *s*-dilated convolution, where $s = 1, 2, 3$, and 4. A dilated filter with dilation factor *s* is a sparse filter of size $(2s + 1) \times (2s + 1)$, where only nine entries of fixed positions are nonzeros.

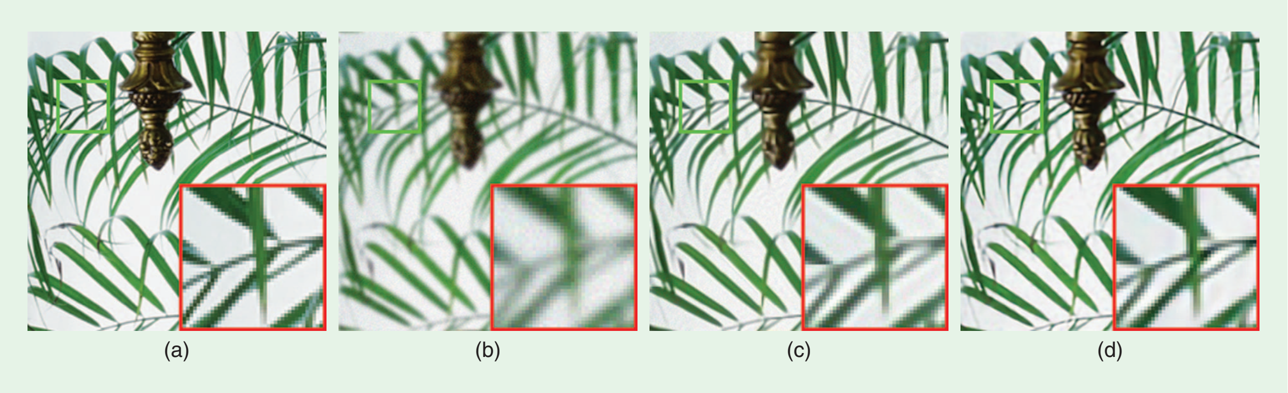

FIGURE 8. Image deblurring results on image *Leaves* (the blur kernel is a Gaussian kernel with standard deviation 1.6 and the noise level is 2). (a) Ground truth, (b) blurred and noisy, (c) NCSR (27.50 dB), and (d) IRCNN (29.78 dB).

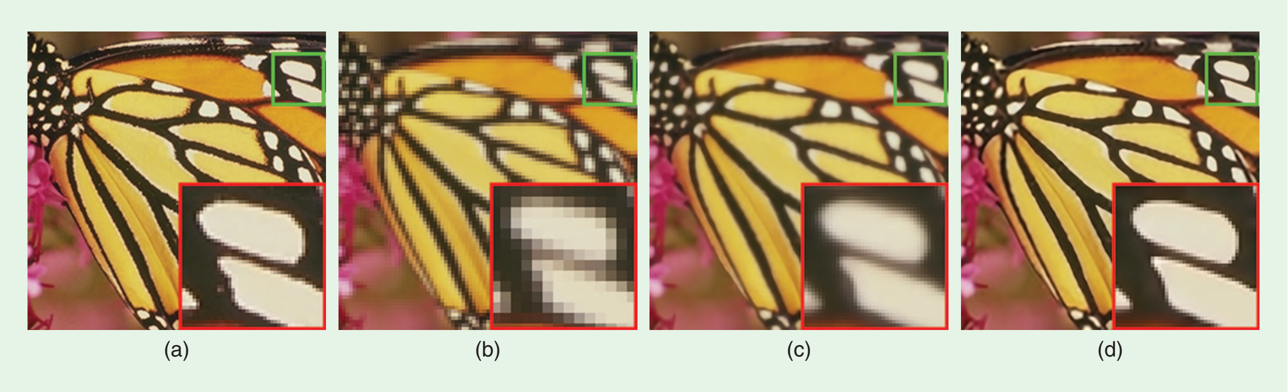

FIGURE 9. SISR results on image *Butterfly* (the blur kernel is a 7 × 7 Gaussian kernel with standard deviation 1.6 and the scaling factor is 3). Note that the comparison with is unfair because it is not retrained for this downsampling setting. This figure is used to show the generality of the proposed method. (a) Ground truth, (b) LR image, (c) VDSR (24.73 dB), and (d) IRCNN (29.32 dB).

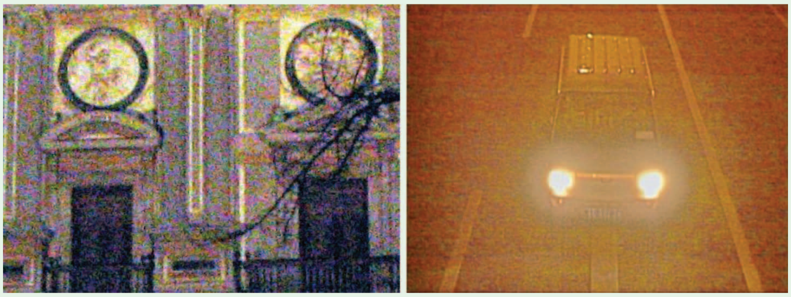

FIGURE 10. Two real-world low-quality images.

IRCNN method can be easily extended to other type of downsampling settings, and achieves very promising result.

Open problems

Though IR has been extensively studied for many years and numerous methods have been developed, it is still a very challenging problem for blind real IR (BRIR). The in-camera pipeline involves many components, including analog-to-digital conversion, white balance, color demosaicking, noise reduction, color transform, tone reproduction, compression, etc., and the quality of a final output image is subjected to many external and internal factors, including illumination, lens, charged-coupled device/complementary metal–oxide–semiconductor sensors, exposure, ISO, camera shaking and object motion, etc. The degradations in real-world images are too complex to be described by simple models such as $\boldsymbol{y} = \boldsymbol{Hx} + \boldsymbol{\upsilon}$. Figure 10 shows two real-world low-quality images, which are hard to be satisfactorily reconstructed by all existing IR methods. The noise therein is strong, non-Gaussian and signal dependent, while the image is LR with nonuniform blur and compression artifacts.

Will deep learning be a good solution to the challenging BRIR problem? We would like to give a positive answer to this question, yet one critical issue is how we can collect the degraded and ground-truth image pairs for training. Note that most of the existing deep CNN-based IR methods are supervised learning methods with simulated training data. AWGN is added to the clean images to simulate the noisy images, and high resolution images are downsampled to simulate the LR images. However, the CNNs trained by such simulated image pairs are much less effective to process the real world degraded images.

How can we train deep models for IR without paired data? Are the recently developed GAN techniques [5] able to tackle this challenging issue? We leave those questions as open problems for future investigations.

What we have learned

We introduced the recent developments of sparse representation, low-rank minimization and deep learning (more specifically deep CNN)-based IR methods. While the image sparsity and low-rankness priors have been dominantly used in past decades, the CNN-based models have been recently rapidly developed to learn deep image priors and have shown promising performance. However, there remain many challenging and interesting problems to investigate for deep learning-based IR. One key issue is the lack of training image pairs in real-world IR applications. It is still an open problem to train deep IR models without using image pairs.

Authors

Lei Zhang (cslzhang@comp.polyu.edu.hk) is a chair professor in the Department of Computing at the Hong Kong Polytechnic University. His research interests include image enhancement and restoration, image quality assessment, image classification, object detection, and visual tracking. He is a Web of Science Highly Cited Researcher selected by Thomson Reuters.

Wangmeng Zuo (wmzuo@hit.edu.cn) is a full professor with the School of Computer Science and Technology at the Harbin Institute of Technology, China. His research interests include image enhancement and restoration, image generation, visual tracking, convolutional network, and image classification. He is a Senior Member of the IEEE.

References

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[2] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image superresolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.

[4] 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.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[6] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang, "Weighted nuclear norm minimization and its applications to low level vision," *Int. J. Comput. Vis.*, vol. 121, no. 2, pp. 183–208, 2017.

[7] J. Kim, J. Kwon Lee, and K. Mu Lee. "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2016, pp. 1646–1654.

[8] C. Ledig, et al. "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[9] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[10] 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. 7, no. 26, pp. 3142–3155, 2017.

[11] K. Zhang, W. Zuo, S. Gu, and L. Zhang. "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

SP