Multimodal Deep Unfolding for Guided Image Super-Resolution

Iman Marivani, Student Member, IEEE, Evaggelia Tsiligianni, Member, IEEE, Bruno Cornelis, Member, IEEE, and Nikos Deligiannis, Member, IEEE

Abstract—The reconstruction of a high resolution image given a low resolution observation is an ill-posed inverse problem in imaging. Deep learning methods rely on training data to learn an end-to-end mapping from a low-resolution input to a high-resolution output. Unlike existing deep multimodal models that do not incorporate domain knowledge about the problem, we propose a multimodal deep learning design that incorporates sparse priors and allows the effective integration of information from another image modality into the network architecture. Our solution relies on a novel deep unfolding operator, performing steps similar to an iterative algorithm for convolutional sparse coding with side information; therefore, the proposed neural network is interpretable by design. The deep unfolding architecture is used as a core component of a multimodal framework for guided image super-resolution. An alternative multimodal design is investigated by employing residual learning to improve the training efficiency. The presented multimodal approach is applied to super-resolution of near-infrared and multi-spectral images as well as depth upsampling using RGB images as side information. Experimental results show that our model outperforms state-of-the-art methods.

Index Terms—multimodal deep unfolding, multimodal image super-resolution, interpretable convolutional neural networks.

I. INTRODUCTION

Image super-resolution (SR) is a well-known inverse problem in imaging, referring to the reconstruction of a high-resolution (HR) image from a low-resolution (LR) observation [1], [2]. The problem is ill-posed as there is no unique mapping from the LR to the HR image. Practical applications such as medical imaging and remote sensing often involve different image modalities capturing the same scene, therefore, another approach in imaging is the joint use of multiple image modalities. The problem of multimodal or guided image super-resolution refers to the reconstruction of an HR image from an LR observation using a guidance image from another modality, also referred to as side information.

Several image processing methods use prior knowledge about the image such as sparse structure [3]–[9] or statistical image priors [10], [11]. Deep learning has been widely used in inverse problems, often outperforming analytical methods [2], [12]. For instance, in single image SR, Convolutional Neural Networks (CNNs) have led to impressive results reported in [13]–[18]. Residual learning [19] enabled the training of very deep neural networks (DNNs), with the models proposed in [20]–[24] achieving state-of-the-art performance.

Capturing the correlation among different image modalities has been addressed with sparsity-based analytical models and coupled dictionary learning [25]–[33]. The main drawback of these approaches is the high-computational cost of iterative algorithms for sparse approximation, which has been addressed by multimodal deep learning methods [34]–[36]. A common approach in multimodal neural network design is the fusion of the input modalities at a shared latent layer, obtained as the concatenation of the latent representations of each modality [34]. The CNN model proposed in [35] for multimodal depth upsampling follows this principle. Nevertheless, current multimodal DNNs are black-box models in the sense that we lack a principled approach to design such models for leveraging the signal structure and properties of the correlation across modalities.

A recent line of research in deep learning for inverse problems considers deep unfolding [37]–[42], that is, the unfolding of an iterative algorithm into the form of a DNN. Inspired by numerical algorithms for sparse coding, deep unfolding designs have been utilized in several imaging problems to incorporate sparse priors into the solution. Results for denoising [42], compressive imaging [43], and image SR [44] have shown that incorporating domain knowledge into the network architecture can improve the performance substantially. Still, these methods focus on single-modal data, thereby lacking a principled way to incorporate knowledge from different imaging modalities. To the best of our knowledge, the only deep unfolding designs for guided image SR have been presented in [45] and [46], that build upon existing unfolding architectures for learned sparse coding [37] and learned convolutional sparse coding [42], respectively.

In this paper, we address the problem of guided image SR with a novel multimodal deep unfolding architecture, which is inspired by a proximal algorithm for convolutional sparse coding with side information. The proximal algorithm is translated into a neural network form coined Learned Multimodal Convolutional Sparse Coding (LMCSC); the network incorporates sparse priors and enables efficient integration of the guidance modality into the solution. While in existing multimodal deep learning methods [34]–[36], it is difficult to understand what the model has learned, our deep neural network is interpretable, in the sense that the model performs steps similar to an iterative algorithm.

The proposed approach builds upon our previous research on multimodal deep unfolding [47], [48], and preliminary results of LMCSC can be found in [49], where the recovery of HR near infrared (NIR) images based on LR NIR observations.
with the aid of RGB images was addressed. In this paper, we integrate LMCSC into different neural network architectures, and present experiments on various multimodal datasets, showing the superior performance of the proposed approach against several single- and multimodal SR methods. Our contribution is as follows:

(i) We formulate the problem of convolutional sparse coding with side information, and propose a proximal algorithm for its solution.

(ii) Inspired by the proposed proximal algorithm, we design a deep unfolding neural network for fast computation of convolutional sparse codes with the aid of side information.

(iii) The deep unfolding operator is used as a core component in a novel multimodal framework for guided image SR that fuses information from two image modalities. Furthermore, we exploit residual learning and introduce skip connections in the proposed framework, obtaining an alternative design that can be trained more efficiently.

(iv) We test our models on several benchmark multimodal datasets, including NIR/RGB, multi-spectral/RGB, depth/RGB, and compare them against various state-of-the-art single-modal and multimodal models. The numerical results show a PSNR gain of up to 2.74 dB over the coupled ISTA method in [45].

The rest of the paper is organized as follows: Section II reviews related work and Section III provides the necessary background. Section IV presents the proposed core deep unfolding architecture for convolutional sparse coding with side information, and our designs for multimodal image SR are presented in Section V followed by experimental results in Section VI. Section VII concludes the paper.

Throughout the paper, all vectors are denoted by boldface lower case letters while lower case letters are used for scalars. We utilize boldface upper case letters to show matrices and boldface upper case letters in math calligraphy to indicate tensors. Moreover, in this paper, the terms upscaling factor and scale are used interchangeably.

II. RELATED WORK

1) Single Image Super-Resolution: A first category of single image SR methods includes interpolation-based methods [11], [50]–[52]. These methods are simple and fast, however, aliasing and blurring effects make them inefficient in obtaining HR images of fine quality. A second category includes reconstruction methods [53]–[55], which use several image priors to regularize the ill-posed reconstruction problem and result in images with fine texture details. Nevertheless, modelling the complex context of natural images with image priors is not an easy task. A third popular category consists of learning-based methods [3]–[8], [13], [18], [20]–[24], which use machine learning techniques to learn the complex mapping between LR and HR images from data.

Among learning based methods, deep learning models have drawn considerable attention as they achieve excellent restoration quality. SRCNN [13] was the first deep learning method for image SR. The model has a simple structure and can directly learn an end-to-end mapping between the LR/HR images. An accelerated version of [13] was presented in [14]. Increasing the depth of CNN architectures ensues several training difficulties which have been mitigated with residual learning. Examples of very deep residual networks for SR include a 20-layer CNN proposed in [15], [16], a 30-layer convolutional autoencoder proposed in [17], and a 52-layer CNN proposed in [22]. Residual learning has also been employed to learn inter-layer dependencies in [23], [24]. An improved residual design obtained by removing unnecessary residual modules was proposed in [20]. Recurrent neural networks (RNNs) have been also used for image SR in [50], [57]. The network in [56] implements a feedback mechanism that carries high-level information back to previous layers, refining low-level encoded information. Following [50], the authors of [57] presented an alternative structure with two states (RNN layers) that operate at different spatial resolutions, providing information flow from LR to HR encodings.

Deep unfolding has been applied to single image SR in [44] where the authors designed a neural network that computes latent representations of the LR/HR image using LISTA [37], a neural network that performs steps similar to the Iterative Soft Thresholding Algorithm (ISTA) [58] (see also Section III).

2) Multimodal Image Super-Resolution: A common approach in multimodal image restoration is the joint or guided filtering approach, that is, the design of a filter that leverages the guidance image as a prior and transfers structural details from the guidance to the target image. Several joint filtering techniques have been proposed in [59]–[61]. Nevertheless, when the local structures in the guidance and target images are not consistent, these techniques may transfer incorrect content to the target image. The approach presented in [62] concerns the design of an explicit mapping that captures the structural discrepancy between images from different modalities.

Model-based techniques and joint filtering methods are limited in characterizing the complex dependency between different modalities. Learning based methods aim to learn this dependency from data. In a depth upsampling method presented in [63], a weighted analysis representation is used to model the complex relationship between depth and RGB images; the model parameters are learned with a task driven training strategy. Another learning based approach relies on sparse modelling and involves coupled-dictionary learning [25]–[32]. Most of these works assume that there is a mapping between the sparse representation of one modality to the sparse representation of another modality. The authors of [33] consider both similarities and disparities between different modalities under the sparse representation invariance assumption.

Purely data-driven solutions for multimodal image SR are provided by multimodal deep learning approaches. Examples include the model presented in [35] implementing a CNN based joint image filter, and the work presented in [36], which is a deep learning reformulation of the widely used guided image filter proposed in [60].

The deep unfolding design LISTA [37] has also been deployed for multimodal image SR in [45] where a coupled ISTA network is presented. The network accepts as input an
LR image from the target modality and an HR image from the
guidance modality. Two LISTA branches are employed
to compute latent representations of the input images. The
estimation of the target HR image is obtained as a linear
combination of these representations. A similar approach is
proposed in [46] that employs three convolutional LISTA
networks to split the common information shared between
modalities, from the unique information belonging to each
modality. The output is then computed as a combination of
these common and unique feature maps after applying the
corresponding dictionaries on them.

III. BACKGROUND

Image super-resolution can be addressed as a linear inverse
problem formulated as follows [3]:

\[ y = Lx + \eta, \]  

(1)

where \( x \in \mathbb{R}^k \) is a vectorized form of the unknown HR
image, and \( y \in \mathbb{R}^n \) is the LR observations contaminated
with noise \( \eta \in \mathbb{R}^n \). The linear operator \( L \in \mathbb{R}^{n \times k}, n < k, \)
describes the observation mechanism, which can be expressed
as the product of a downsampling operator \( E \) and a blurring
filter \( H \) [3]. Problem (1) appears in many imaging applications
including image restoration and inpainting [1, 2].

A. Image Super-Resolution via Sparse Approximation

Even when the linear observation operator \( L \) is given, problem (1) is ill-posed and requires additional regularization
for its solution. Sparsity has been widely used as a regularizer
leading to the well-known sparse approximation problem [64].
Instead of directly solving for \( x \), in this paper, we rely on
a sparse modelling approach presented in [3]. According to [3],
an \( n \)-dimensional (vectorized) patch \( y \) from a bicubic-upscaled
LR image and the corresponding patch \( x \) from the respective
HR image can be expressed by joint sparse representations.
By jointly learning two dictionaries \( \Psi_y \in \mathbb{R}^{n \times m}, \Psi_x \in \mathbb{R}^{n \times m}, \)
\( n \leq m \), for the low- and the high-resolution image patches,
respectively, we can enforce the similarity of sparse representations
of patch pairs such that \( y = \Psi_y \alpha \) and \( x = \Psi_x \alpha \),
\( \alpha \in \mathbb{R}^m \). Then, computing the HR patch \( x \) is equivalent to
finding the sparse representation of the LR patch \( y \), by solving

\[ \min_{\alpha} \frac{1}{2} \| y - \Psi_y \alpha \|_2^2 + \lambda \| \alpha \|_1, \]  

(2)

where \( \lambda \) is a regularization parameter, and \( \| \alpha \|_1 = \sum_{i=1}^{k} |\alpha_i| \)
is the \( \ell_1 \)-norm, which promotes sparsity. Several methods have
been proposed for solving (2) including pivoting algorithms,
interior-point methods and gradient based methods [64].

Sparse modelling techniques that involve computations applied
to independent image patches do not take into account the
consistency of pixels in overlapping patches [9]. Convolutional
Sparse Coding (CSC) [65] is an alternative approach, which can
be directly applied to the entire image. Denoting with \( Y \in \mathbb{R}^{n_1 \times n_2} \) the image of interest, the convolutional
sparse codes are obtained by solving the following problem:

\[ \min_{U_i} \frac{1}{2} \| Y - \sum_{i=1}^{k} D_i \ast U_i \|_2^2 + \lambda \sum_{i=1}^{k} \| U_i \|_1, \]  

(3)

where \( \| A \|_F = \sqrt{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} |a_{ij}|^2} \) is the Frobenius norm,
\( D_i \in \mathbb{R}^{n_1 \times p_2}, i = 1, \ldots, k \), are the atoms of a convolutional
dictionary \( \mathcal{D} \in \mathbb{R}^{n_1 \times p_2 \times k} \), and \( U_i \in \mathbb{R}^{n_1 \times n_2}, i = 1, \ldots, k \),
are the sparse feature maps with respect to \( \mathcal{D} \). The \( \ell_1 \)-norm computes the sum of absolute values of the elements in \( U_i \)
(as if \( U_i \) is unrolled as a vector). Efficient solutions of (3) are
presented in [66, 67].

According to recent studies [68], the accuracy of sparse
approximation problems can be improved if a signal \( \omega \) correlated
with the target signal \( y \) is available; we refer to \( \omega \)
as side information (SI). Assume that \( y \in \mathbb{R}^n \) and \( \omega \in \mathbb{R}^d \)
have similar sparse representations \( \alpha \in \mathbb{R}^m, s \in \mathbb{R}^m \), under
dictionaries \( \Psi_y \in \mathbb{R}^{n \times m}, \Psi_\omega \in \mathbb{R}^{d \times m}, n \leq m, d \leq m \),
respectively. Then the sparse representation \( \alpha \) can be obtained
as the solution of the \( \ell_1-\ell_1 \) minimization problem [68].

\[ \min_{\alpha} \frac{1}{2} \| y - \Psi_y \alpha \|_2^2 + \lambda (\| \alpha \|_1 + \| \alpha - s \|_1). \]  

(4)

Problem (4) has been theoretically studied in [68]. Numerical
methods for its solution are presented in [69, 70].

B. Deep Unfolding

Analytical approaches for sparse approximation are usually
equipped with theoretical guarantees; however, their major
drawback is their high computational complexity. In some
applications, the deployed dictionaries also need to be learned,
increasing the computational burden [3]. The authors of [37]
address this problem by a neural network design that performs
operations similar to the Iterative Soft Thresholding Algorithm
(ISTA) [58] proposed for the solution of (4). The learning
process results in a trained version of ISTA, coined LISTA.
The \( t \)-th layer of LISTA computes:

\[ \alpha^t = \phi_\gamma(S\alpha^{t-1} + W_y), \quad \alpha^0 = 0, \]  

(5)

where

\[ \phi_\gamma(v_i) = \text{sign}(v_i) \max\{0, |v_i| - \gamma\}, \quad i = 1, \ldots, k, \]  

(6)

is the soft thresholding operator; \( S \in \mathbb{R}^{m \times m}, W \in \mathbb{R}^{m \times n} \)
and \( \gamma > 0 \) are parameters, which are fixed in ISTA, while
LISTA learns them from data. As a result, LISTA achieves
high accuracy in only a few iterations. The technique known as
deep unfolding was also explored in [38]–[41]; a convolutional
LISTA design for CSC was presented in [42].

The aforementioned deep unfolding studies deal with single-modal
data. A deep unfolding design that incorporates side
information coming from another modality was first presented
in our previous work [47]. The model proposed in [47] relies
on a proximal method for the solution of (4), which iterates
over

\[ \alpha^t = \xi_\mu(I - \frac{1}{L} \Psi^T \Psi)^{\alpha^{t-1}} + \frac{1}{L} \Psi^T y; s), \quad \alpha^0 = 0, \]  

(7)

with \( \mu, L \) appropriate parameters. The proximal operator
\( \xi_\mu \) incorporates the side information \( s \) and is expressed as follows:
Fig. 1. The proposed LMCSC model with unfolded recurrent stages. The model computes sparse feature maps $\mathbf{U}$ of an image $\mathbf{Y}$ given sparse feature maps $\mathbf{Z}$ of the side information. The nonlinear activation function follows the proximal operator $\xi_{\mu}(v; z)$ given by (8).

1) For $s_i \geq 0$, $i = 1, \ldots, m$:

$$
\xi_{\mu}(v_i; s_i) = \begin{cases} 
    v_i + 2\mu, & v_i < -2\mu \\
    0, & -2\mu \leq v_i \leq 0 \\
    v_i, & 0 < v_i < s_i \\
    s_i, & s_i \leq v_i \leq s_i + 2\mu \\
    v_i - 2\mu, & v_i \geq s_i + 2\mu 
\end{cases} 
$$

(8)

2) For $s_i < 0$, $i = 1, \ldots, m$:

$$
\xi_{\mu}(v_i; s_i) = \begin{cases} 
    v_i + 2\mu, & v_i < s_i - 2\mu \\
    s_i - 2\mu \leq v_i \leq s_i \\
    v_i, & s_i < v_i < 0 \\
    0, & 0 \leq v_i \leq 2\mu \\
    v_i - 2\mu, & v_i \geq 2\mu 
\end{cases} 
$$

(9)

By writing the proximal algorithm in the form

$$
\mathbf{y}^t = \xi_{\mu}(S\mathbf{y}^{t-1} + \mathbf{W}u; s), \quad \mathbf{y}^0 = 0, 
$$

(10)

and translating (10) into a deep neural network, we obtain a fast multistage operator referred to as Learned Side-Information-driven iterative soft Thresholding Algorithm (LeSITA). LeSITA has a similar expression to LISTA (5), however, (10) employs the new activation function $\xi_{\mu}$ that integrates side information into the learning process.

IV. DESIGN MULTIMODAL CONVOLUTIONAL NETWORKS WITH DEEP UNFOLDING

In what follows, we consider that, besides the observations $\mathbf{Y}$ of the target signal, another image modality $\mathbf{\Omega}$, correlated with $\mathbf{Y}$ is available. We assume that the two image modalities can be represented by convolutional sparse codes that are similar by means of the $\ell_1$-norm. Specifically, let $\mathbf{Y} = \sum_{i=1}^{k} D_{i}^Y \ast \mathbf{U}_{i}$ be a sparse representation of the observed image $\mathbf{Y}$ with respect to a convolutional dictionary $\mathbf{D}^Y$. $D_{i}^Y \in \mathbb{R}^{n_1 \times n_2}$, $i = 1, \ldots, k$, denote the atoms of $\mathbf{D}^Y$. By employing a convolutional dictionary $\mathbf{D}^\Omega$ with atoms $D_{i}^\Omega \in \mathbb{R}^{p_1 \times p_2}$, $i = 1, \ldots, k$, the guidance image $\mathbf{\Omega} \in \mathbb{R}^{n_1 \times n_2}$ can be expressed as $\mathbf{\Omega} = \sum_{i=1}^{k} D_{i}^\Omega \ast \mathbf{Z}_{i}$, with the convolutional sparse codes $\mathbf{Z}_{i}$, $i = 1, \ldots, k$, obtained as the solution of (3). Then, we can compute the unknown sparse codes $\mathbf{U}_{i}$ of the target modality by solving a problem formulated in a way similar to (3), that is,

$$
\min_{\mathbf{U}_{i}} \frac{1}{2} \| \mathbf{Y} - \sum_{i=1}^{k} D_{i}^Y \ast \mathbf{U}_{i} \|_F^2 + \lambda \sum_{i=1}^{k} \| \mathbf{U}_{i} \|_1 + \sum_{i=1}^{k} \| \mathbf{U}_{i} - \mathbf{Z}_{i} \|_1. 
$$

(11)

There is a correspondence between convolutional and linear sparse codes. If we replace the convolutional dictionary $\mathbf{D}^Y$ with a matrix $\mathbf{A}$ with Toeplitz structure, and take into account the linear properties of convolution, then (11) reduces to (3). Specifically, we define $\mathbf{A} \in \mathbb{R}^{(n_1 - p_1 + 1)(n_2 - p_2 + 1) \times kn_1n_2}$ as a sparse dictionary obtained by concatenating the Toeplitz matrices that unroll $\mathbf{D}_{i}^Y \ast \mathbf{s}_i$ in $\mathbb{R}^{kn_1n_2}$ and $\mathbf{s} \in \mathbb{R}^{kn_1n_2}$ take the form of vectorized sparse feature maps of the target and the side information images, respectively. Then, by replacing the convolutional operations in (11) with multiplications, we obtain (4), and the proximal algorithm (7) can be employed to compute convolutional sparse codes.

Nevertheless, transforming (11) to (4) and using (7) for its solution is not computationally efficient. Since CSC deals with the entire image, the dimensionality of (3) becomes too high and the proximal method becomes impractical. We use the correspondence between linear and convolutional representations, and formulate an iterative algorithm that performs convolutions as follows: In the proximal algorithm (7), the matrices $\mathbf{A}$ and $\mathbf{A}^T$, which take the form of concatenated Toeplitz matrices in the convolutional case, are replaced by the convolutional dictionaries $\mathbf{B}^Y$ and $\mathbf{B}^\Omega$, respectively. Then, by replacing multiplications with convolutional operations, we can compute the convolutional codes $\mathbf{U}$ of the target image, given the convolutional codes $\mathbf{Z}$ of the guidance image, by iterating over:

$$
\mathbf{U}^t = \xi_{\mu}(\mathbf{U}^{t-1} - \mathbf{B}^Y \ast \mathbf{B}^{\Omega} \ast \mathbf{U}^{t-1} + \mathbf{B}^Y \ast \mathbf{Y}; \mathbf{Z}), 
$$

(12)

where $\mathbf{U}$, $\mathbf{Z}$ are tensors of size $p_1 \times p_2 \times k$.

Equation (12) can be translated into a deep convolutional neural network (CNN). Each stage of the network computes the sparse feature maps according to

$$
\mathbf{U}^t = \xi_{\mu}(\mathbf{U}^{t-1} - \mathbf{Q} \ast \mathbf{R} \ast \mathbf{U}^{t-1} + \mathbf{P} \ast \mathbf{Y}; \mathbf{Z}), 
$$

(13)

with $\mathbf{Q} \in \mathbb{R}^{p_1 \times p_2 \times c \times k}$, $\mathbf{R} \in \mathbb{R}^{p_1 \times p_2 \times k \times c}$, $\mathbf{P} \in \mathbb{R}^{p_1 \times p_2 \times c \times k}$ learnable convolutional layers and $\mu > 0$ a learnable parameter; $c$ is the number of channels of the employed images. The proposed network architecture, depicted in Fig. 1, is referred to as Learned Multimodal Convolutional Sparse Coding (LMCSC). The network can be trained in a supervised manner to map an input image to sparse feature maps. During training, the parameters $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{P}$ and $\mu$ are learned; therefore, the deep LMCSC can achieve high accuracy with only a fraction of computations of the proximal method.

LMCSC uses the convolutional sparse codes $\mathbf{Z}$ of the guidance modality $\mathbf{\Omega}$ to compute the convolutional sparse...
codes of the target modality $Y$. An efficient multimodal convolutional operator should integrate a fast operator for the encoding of the guidance modality. In the models presented next, we obtain $\mathcal{Z}$ using the ACSC operator presented in [42]. ACSC has the form of convolutional LISTA, with the $t$-th layer computing:

$$\mathcal{Z}^t = \phi_t(\mathcal{Z}^{t-1} - \mathcal{T} \ast \mathcal{V} \ast \mathcal{Z}^{t-1} + \mathcal{G} \ast \Omega),$$

(14)

where $\phi_t$ is the proximal operator given by (6). The parameters of the convolutional layers $\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times c \times k}$, $\mathcal{G} \in \mathbb{R}^{p_1 \times p_2 \times c \times k}$ and $\mathcal{V} \in \mathbb{R}^{p_1 \times p_2 \times c \times c}$, are learned from data. The architecture of ACSC is depicted in Fig. [2].

V. DEEP MULTIMODAL IMAGE SUPER-RESOLUTION

The proposed LMCSC architecture can be employed to perform multimodal image super-resolution based on a sparsity-driven convolutional model. The proposed model follows similar principles with [3]. Specifically, the sparse linear modelling of LR/HR image patches presented in [3] is replaced by a sparse convolutional modelling of the entire LR/HR images, followed by similarity assumptions between the convolutional representations of the LR and HR images. To efficiently integrate information from a second image modality, we also assume that the target and the guidance image modalities are similar by means of the $\ell_1$-norm in the representation domain.

A. LMCSC-Net

Our first model proposed for multimodal image super-resolution relies on the following assumption: The LR observation $Y$ and the HR image $X$ share the same convolutional sparse features maps $U_i$, under different convolutional dictionaries $D^Y$ and $D^X$, that is, $Y = \sum_{i=1}^{k} D^Y_i \ast U_i$, $X = \sum_{i=1}^{k} D^X_i \ast U_i$, where $D^Y_i$, $D^X_i$ are the atoms of the respective convolutional dictionaries. Given $D^Y$, $D^X$, finding a mapping from $Y$ to $X$ is equivalent to computing the sparse features maps $\hat{U}$ of the observed LR image $Y$. The similarity assumption between the target and the guidance image modalities in the representation domain implies that the convolutional sparse codes $\mathcal{Z}$ of the guidance HR image $\Omega$ are similar to $\hat{U}$ by means of the $\ell_1$-norm. Therefore, $\hat{U}$ can be obtained as the solution of the $\ell_1$-$\ell_1$ minimization problem (11). Based on these assumptions, we build our first model for multimodal image SR using LMCSC as a core component of a deep architecture.

The proposed model, coined LMCSC-Net, consists of three subnetworks: (i) an LMCSC encoder that produces convolutional latent representations of the input LR image with the aid of side information, (ii) a side information encoder that produces latent representations of the guidance HR image, and (iii) a convolutional decoder that computes the target HR image. The goal of the LMCSC encoder is to learn a convolutional sparse feature map $U$ of the LR image $Y$, also shared by the HR image $X$, using a convolutional sparse feature map $\mathcal{Z}$ as side information, akin to the model presented in Section IV. The LMCSC branch is followed by a convolutional decoder realized by a learnable convolutional dictionary $D^X$.

The decoder receives the latent representations $U$ provided by LMCSC and estimates $X$ according to $\hat{X} = D^X \ast \hat{U}$. The entire network, depicted in Fig. [3] is trained end-to-end using the mean square error (MSE) loss function:

$$\min_{\Theta} \sum_{i} \| \hat{X}_i - X_i \|^2_F,$$

(15)

where $\Theta$ denotes the set of all network parameters, $X_i$ is the ground-truth image of the target modality, and $\hat{X}_i$ is the estimation computed by the network.

Different from our previous multimodal image SR design [48] which relies on LeSITA [47], LMCSC-Net has a novel convolutional structure inspired by a different proximal algorithm. The core LMCSC component computes latent representations of the target modality using side information from the guidance modality, performing fusion of information at every layer. Therefore, our approach is different from coupled ISTA [45] which employs one branch of LISTA [37] for each modality and fuses the latent representations only in the last layer.

B. LMCSC$_+$-Net

The model presented in Section V-A learns similar sparse representations of three different image modalities, that is, the input LR image $Y$, the guidance modality $\Omega$ and the HR image $X$. Learning representations that mainly encode the common information among the different modalities is critical for the performance of the model. Nevertheless, some information from the guidance modality may be misleading when learning a mapping between $X$ and $Y$. In other words, the encoding performed by the ACSC branch may result in transferring unrelated information to the LMCSC encoder. As a result, the latent representation of the target modality may
not capture the underlying mapping between the LR and HR images in the representation domain. Furthermore, assuming identical latent representations for both LR and HR images limits the performance of the network especially when the degradation level of the LR observations is high.

In order to address the aforementioned problems, we relax the assumption concerning the similarity between the LR and HR images in the representation domain. Specifically, we assume that the convolutional sparse codes \( U \) of the HR image \( X \) can be obtained as a non-linear transformation of the respective codes \( U \) of the LR image \( Y \), that is, \( U = F_y(U_y) \) where \( F_y \) is a non-linear function parameterized by \( \Theta \).

Under this assumption, we build the proposed multimodal SR framework by employing the following components: (i) An LMCSC subnetwork is used to fuse the information from the LR observations and the guidance HR image, providing a first estimation of the target HR image with the aid of side information. (ii) An ACSC subnetwork following the LMCSC subnetwork is used to enhance the transformation between the LR and HR images of the target modality without using side information. The architecture of the proposed model, referred to as LMCSC+Net, is depicted in Fig. 4. The additional ACSC and dictionary layers in LMCSC+Net implement \( F_y \). The network is trained using the objective (15).

C. LMCSC-Net with Skip Connections

Considering the significant improvement achieved by residual learning in the training efficiency and the prediction accuracy [19]–[24], we enhance the proposed LMCSC+Net with a skip connection, introducing a new model coined LMCSC-ResNet. For the design of LMCSC-ResNet we rely on the assumption that the HR image \( X \) contains all the low-frequency information from the LR image \( Y \) plus some high-frequency details that can be captured by a non-linear mapping between LR and HR images. By using an identity mapping of the input, the capacity of the network can be assigned to learning the high frequency details, since the low-frequency information is provided by \( Y \). LMCSC-ResNet learns the non-linear mapping \( H(Y, \Omega) = F(Y, \Omega) + Y \), where \( F(Y, \Omega) = X - Y \). In LMCSC-ResNet, \( F(Y, \Omega) \) is obtained from the LMCSC+Net. The model architecture is presented in Fig. 5. This model is also trained end-to-end using the objective (15).

VI. EXPERIMENTS

We apply the proposed models to different upsampling tasks, that is, super-resolution of near-infrared (NIR) images, depth upsampling and super-resolution of multi-spectral data, using RGB images as side information. We compare our models against state-of-the-art single-modal and multimodal methods showing the superior performance of the proposed approach. Before demonstrating our experimental results, we present the employed datasets and report implementation details.

A. Datasets

1) EPFL RGB-NIR dataset: NIR images are acquired at a low resolution due to the high cost per pixel of a NIR sensor compared to an RGB sensor. We employ the EPFL RGB-NIR dataset\(^1\) and apply our models to super-resolve an LR NIR image with the aid of an HR RGB image. The dataset contains 477 spatially aligned NIR/RGB image pairs. Our training set contains approximately 30,000 cropped image pairs extracted from 50 images. Each training image is of size 44 x 44 pixels; the size is chosen with respect to memory requirements and computational complexity. We also create a testing set containing 25 image pairs; testing is performed on an entire image.

The NIR images consist of one channel. An LR version of a NIR image is generated by blurring and downsampling the ground truth HR version. We convert the RGB images to YCbCr and only utilize the luminance channel as the side information. Following [13], we apply bicubic interpolation as a pre-processing step to upscale the LR input such that the input and output images are of the same size.

2) NYU v2 RGB-D dataset: Depth cameras like Microsoft Kinect and time-of-flight (ToF) cameras only provide low-resolution depth images. Therefore, depth upsampling is a necessary task for many vision applications. We apply our models for depth upsampling with the aid of RGB images, using the NYU v2 RGB-D dataset [71]. The dataset contains 1449 RGB images with their depth maps. Similar to [35], we use the first 1000 images for training, and the remaining 449 for testing.

3) Columbia multi-spectral database: The third dataset that we use to evaluate the proposed models is the Columbia multispectral database\(^2\) which contains spectral reflectance data and RGB images. For testing, we reserve 7 images from the 640 nm band and randomly select 7 images from different bands; the rest are used for training.

B. Implementation Details

All networks are designed with three unfolding steps for the target (LMCSC) and the side information (ACSC) encoders. The number of unfolding steps is chosen after taking into account the trade-off between the computational complexity and the reconstruction accuracy; for instance, by increasing the unfolding steps to five, the improvement in the average PSNR

\(^1\)https://ivrl.epfl.ch/supplementary_material/cvpr11/
\(^2\)http://www.cs.columbia.edu/CAVE/databases/multispectral
TABLE I
SUPER-RESOLUTION OF NIR IMAGES WITH THE AID OF RGB IMAGES. PERFORMANCE COMPARISON [IN TERMS OF AVERAGE PSNR (dB)] OVER ALL TEST IMAGES FOR ×2, ×4 AND ×6 UPSCALING.

| Scale | CSCN [14] | ACSC [12] | EDSR [22] | SRFB [29] | DJF [35] | CoISTA [45] | LMCSC-Net | LMCSC-Net | LMCSC-ResNet |
|-------|-----------|-----------|-----------|-----------|---------|------------|------------|------------|-------------|
| ×2    | 36.84     | 36.92     | 36.99     | 36.94     | 38.78   | 38.93      | 39.28      | 39.42      | 39.57       |
| ×4    | 30.64     | 30.35     | 31.85     | 32.11     | 32.91   | 33.20       | 33.96      | 34.23      | 34.36       |
| ×6    | 27.94     | 27.91     | 30.49     | 30.63     | 29.19   | 30.88       | 32.07      | 31.94      | 32.33       |

TABLE II
SUPER-RESOLUTION OF NIR IMAGES WITH THE AID OF RGB IMAGES. PERFORMANCE COMPARISON [IN TERMS OF PSNR (dB) AND SSIM] FOR SELECTED TEST IMAGES FOR ×2, ×4 AND ×6 UPSCALING.

| NIR/RGB | u-0004 | u-0006 | u-0017 | u-0018 | u-0020 | u-0026 | u-0030 | u-0050 | Average |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| ×2      |        |        |        |        |        |        |        |        |         |
| PSNR    | 30.55  | 30.92  | 36.89  | 0.946  | 34.83  | 0.918  | 30.50  | 0.911  | 32.68  |
| SSIM    | 30.51  | 0.914  | 30.89  | 0.917  | 30.58  | 0.910  | 30.51  | 0.916  | 32.19  |
| ×4      |        |        |        |        |        |        |        |        |         |
| PSNR    | 28.70  | 0.887  | 36.99  | 0.945  | 34.83  | 0.918  | 30.50  | 0.911  | 32.68  |
| SSIM    | 30.51  | 0.914  | 30.89  | 0.917  | 30.58  | 0.910  | 30.51  | 0.916  | 32.19  |
| ×6      |        |        |        |        |        |        |        |        |         |
| PSNR    | 26.97  | 0.878  | 36.99  | 0.945  | 34.83  | 0.918  | 30.50  | 0.911  | 32.68  |
| SSIM    | 30.51  | 0.914  | 30.89  | 0.917  | 30.58  | 0.910  | 30.51  | 0.916  | 32.19  |

is less than 0.1 dB while the execution time is almost 87% higher. In the LMCSC-ResNet, the ACSC branch employed for the nonlinear mapping of the target signal is designed with one unfolding step.

We empirically set the size of the network parameters $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{r}$ and $\mathcal{T}$ to $7 \times 7 \times 1 \times 85$; the size of $\mathcal{R}$ and $\mathcal{V}$ are set to $7 \times 7 \times 85$. The size of the convolutional dictionaries for reconstruction is $7 \times 7 \times 85 \times 1$. Note that a convolutional layer of size $k \times k \times c \times c$ consists of $g$ convolutional filters with kernel size $k$ and $c$ channels. We use untied weights at every unfolding step, i.e., the $t$-th layer of LMCSC and ACSC subnetworks is realized by the independent variables $\mathcal{R}_t$, $\mathcal{Q}_t$, $\mathcal{V}_t$, $\mathcal{T}_t$, respectively. The weights of all layers are initialized randomly using the Gaussian distribution with
standard deviation equal to 0.01. The parameters $\mu$ and $\gamma$ of the proximal operators are both initialized to 0.2. We train the network using the Adam optimizer.

We notice that the complexity of our networks is dominated by the LeSIFT activation layers in the LMCSC block, and an implementation based on [6] and [7] is not efficient. In order to address this issue, we rewrite the proximal operator in (8), (9) as follows:

$$
\xi_\mu(v_i; s_i) = \text{sign}(s_i)[R(\text{sign}(s_i)v_i - 2\mu - |s_i|) - R(\text{sign}(s_i)v_i - |s_i|) + R(\text{sign}(s_i)v_i) - R(-\text{sign}(s_i)v_i - 2\mu)],
$$

where $R(v) = \max(0, v)$ is the Rectified Linear Unit (ReLU) function. This form of the proximal operator results in a 30% faster implementation and we use this version in all of the experiments.

C. Performance Comparison

1) Super-resolution of NIR/RGB images: The first set of experiments involves super-resolution of NIR images with the aid of RGB images. The proposed models are compared with several single-modal and multimodal methods. Among the reference single-modal methods, the network designs proposed in [44] and [42] also involve deep unfolding architectures with sparse priors. [20] is a residual network, while [56] has an RNN structure. The multimodal designs include [35] and [45]; recall that the latter is a LISTA-based network.

We employ the EPFL RGB-NIR dataset and test our models on 25 NIR/RGB image pairs. Table IV presents average results in terms of the Peak Signal-to-Noise Ratio (PSNR) for upsampling factors equal to $\times 2$, $\times 4$, and $\times 6$. We also present detailed PSNR and structural similarity (SSIM) index results for selected test images. In Table IV, these images have been used for testing in [33]. As can be seen in Table IV, our models deliver higher reconstruction accuracy compared to reference methods at all scales, with LMCSC-ResNet achieving the best performance. Furthermore, the numerical results show that as the upsampling factor increases, the PSNR gain of LMCSC-ResNet over the second best reference method [45] also increases. For instance, at scale $\times 2$ the gain is 0.64 dB, and rises to 0.84 dB and 1.45 dB for scales $\times 4$ and $\times 6$, respectively. The results for selected images in Table IV are similar. The average PSNR gain over [45] is 1.7 dB at scale $\times 2$, and grows to 1.7 dB at scale $\times 4$, and 2.03 dB at scale $\times 6$. The role of multimodal fusion becomes more significant as the SR upsampling factor increases; our numerical analysis shows that the proposed models can effectively fuse the information from two different modalities.

2) Depth upsampling: For the application of depth map upsampling with the aid of RGB images, we train our LMCSC-ResNet network for three upsampling factors, $\times 4$, $\times 8$, and $\times 16$, using the first 1000 images of the NYU v2 dataset [71]. We report averaged Root Mean Square Error (RMSE) results over 449 test images in Table III and also detailed results for 10 selected images from this dataset in Table IV. Table III involves comparison with several multimodal methods such as the joint filtering approaches JBF [59], GF [60], SDF [61], a learning based method proposed by Gu et al. [63], and the deep learning designs DJF [35]; the numerical results for the reference methods are provided by the authors of [63]. We select the second and third best methods from Table III that is, [35] and [63], to compare the performance of our models on selected images in Table IV. We note that for the scales $\times 4$ and $\times 8$, LMCSC-ResNet achieves the lowest average RMSE. However, for an upsampling factor $\times 16$ LMCSC-ResNet has the best performance. At the highest upsampling factor, in the presence of very limited information from the LR input, LMCSC-ResNet has an RMSE gain of 0.36 over the second best method, i.e., Gu et al. [63].

3) Multi-spectral data super-resolution: We utilize the Columbia multi-spectral dataset for the last set of experiments. We apply LMCSC-ResNet to super-resolve spectral images using their RGB version as side information. The experiments involve comparison of our models against several multimodal methods, namely, JBF [59], GF [60] and SDF [61], which are joint filtering approaches, DJF [35] and CoISTA [45] which are deep learning designs, the method proposed in [62] coined JFSM, and the coupled dictionary learning based method CDLSR [33]. We also report results for the single-modal deep learning designs EDSR [20] and SRFBN [56]. Numerical results are presented in Tables V and VI. Table V includes results for $\times 4$ upsampling in terms of PSNR and SSIM for 7 test images of the 640nm band. We can see that LMCSC-ResNet yields the best performance for scale $\times 4$ with a gain of more than 2.1 dB over [45]. Table VI presents detailed results for three different scales, $\times 4$, $\times 8$ and $\times 16$, for randomly selected test images from different bands. Similar to the depth map upsampling application, Table VI shows that for $\times 4$ and $\times 8$ scales LMCSC-Net provides the highest average PSNR. For $\times 16$, LMCSC-ResNet delivers the best reconstruction accuracy with a PSNR gain of 2.74 dB over [45], the second best method among the competing works.

4) Visual examples: We conclude our experiments with visual examples for two super-resolved NIR images, presented in Fig. 7 and Fig. 6, as well as for two super-resolved multi-spectral images, presented in Fig. 8 and Fig. 9. The visual comparison corroborates the presented numerical results.

D. Complexity Analysis

The proposed LMCSC-based networks operate on the entire image rather than reconstructing overlapping patches and aggregating them. The inference time of the proposed networks is independent of the upsampling factor, as the input is first upsampled to the desired resolution using bicubic interpolation. Table VII reports averaged inference running times of the proposed models and CoISTA [35] on 10 test images from the Columbia multi-spectral database with size $512 \times 512$ pixels and on 10 images of size $256 \times 256$ pixels, which are cropped versions of the previous ones. All models are
TABLE III

DEPTH UPSAMPLING WITH THE AID OF RGB IMAGES: PERFORMANCE COMPARISON [IN TERMS OF AVERAGE RMSE] OVER 449 TEST IMAGES FROM THE NYU v2 RGB-D DATASET FOR ×4, ×8, AND ×16 UPSAMPLING.

| RGB-D | GF [60] | JBF [59] | SDF [61] | DJF [35] | Gu et al. [63] | LMCSC-Net | LMCSC+Net | LMCSC-ResNet |
|-------|---------|---------|---------|---------|-------------|-----------|-----------|-------------|
| ×4    | 4.04    | 2.31    | 3.04    | 1.97    | 1.56        | 1.49      | 1.38      | 1.45        |
| ×8    | 7.34    | 4.12    | 5.67    | 3.39    | 2.99        | 2.67      | 2.58      | 2.61        |
| ×16   | 12.23   | 6.98    | 9.97    | 5.63    | 5.24        | 5.01      | 4.93      | 4.88        |

Fig. 6. Super-resolution of the NIR image “i-0013” with upscaling factor ×4. (a) ground-truth (b) CSCN [44], (c) ACSC [42], (d) EDSR [20], (e) DJF [35] and (f) LMCSC-ResNet.

Fig. 7. Super-resolution of the test NIR image “u-0000” with upscaling factor ×4. (a) ground-truth (b) CSCN [44], (c) ACSC [42], (d) EDSR [20], (e) DJF [35] and (f) LMCSC-ResNet. Results for CDLSR [33] are not presented as the code is not available.
tested on an NVIDIA GeForce GTX 1070. We observe that our proposed networks have higher inference times compared to CoISTA [45], which is due to the proposed fusion strategy. Specifically, CoISTA [45] fuses modalities in the last layer of the network by a linear combination of the sparse representations. In contrast, the LMCSC layers of our networks perform intermediate fusion using the activation function defined in [8] and [9]. As demonstrated by the reconstruction accuracy results, where gains of up to 2.74dB over CoISTA [45] were reported, the proposed fusion approach can effectively capture the complex relationships across modalities at the expense of a reasonable increase in terms of computation.

VII. CONCLUSIONS

In this paper, we presented a new approach for guided image super-resolution based on a novel multimodal deep unfolding design. The efficient integration of the guidance modality into the deep learning architecture is achieved with a neural network that performs steps similar to an iterative algorithm for convolutional sparse coding with side information. By exploiting residual learning, we further improve the training efficiency and increase reconstruction accuracy of the proposed framework. Our approach was applied to super-resolution of NIR images, multi-spectral and depth maps with the aid of HR RGB images. The superior performance of our models against various state-of-the-art single-modal and multimodal methods was demonstrated by experimental results.

REFERENCES

[1] A. Ribes and F. Schmitt, “Linear inverse problems in imaging,” *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 84–99, 2008.
[2] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsevgalos, “Using deep neural networks for inverse problems in imaging: Beyond analytical methods,” *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 20–36, 2018.
[3] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” *IEEE Transactions on Image Processing*, vol. 19, pp. 2861–2873, 2010.
[4] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. S. Huang, “Coupled dictionary training for image super-resolution,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, 2012.
[5] R. Timofte, V. D. Smet, and L. V. Gool, “Anchored neighborhood regression for fast example-based super-resolution,” in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
TABLE VI
Super-resolution of multi-spectral images with the aid of RGB images. Performance comparison [in terms of PSNR (dB) and SSIM] over selected multi-spectral test images (from different bands) for x4, x8 and x16 upscaling.

| MS/RGB | Chart toy | Egyptian | Feathers | Glass tiles | Jelly beans | Oil Paintings | Paints | Average |
|--------|-----------|----------|----------|-------------|-------------|---------------|--------|---------|
| ×4     | PSNR      | SSIM     | PSNR     | SSIM        | PSNR        | SSIM          | PSNR   | SSIM    |
| Bicubic| 28.94     | 0.9424   | 30.57    | 0.9780      | 30.80       | 0.9562        | 26.65  | 0.9242  |
| SDF [6]| 31.87     | 0.9094   | 34.43    | 0.9795      | 33.45       | 0.9650        | 30.22  | 0.9374  |
| JBF [5]| 32.56     | 0.9651   | 38.73    | 0.9775      | 33.60       | 0.9687        | 27.52  | 0.9341  |
| JFSM [6]| 32.98     | 0.9295   | 40.39    | 0.9705      | 33.89       | 0.9425        | 28.98  | 0.9397  |
| GF [6] | 34.09     | 0.9788   | 40.24    | 0.9796      | 33.60       | 0.9748        | 29.46  | 0.9593  |
| EDSR [20]| 33.45   | 0.9836   | 40.03    | 0.9829      | 35.55       | 0.9875        | 29.75  | 0.9736  |
| SRBFIN [56]| 33.93  | 0.9838   | 40.04    | 0.9822      | 35.53       | 0.9873        | 29.53  | 0.9676  |
| DGE [5] | 34.19     | 0.9559   | 37.81    | 0.9620      | 32.11        | 0.9336        | 28.94  | 0.9459  |
| DIF [55]   | 37.86     | 0.9935   | 45.69    | 0.9922      | 40.13        | 0.9939        | 34.07  | 0.9915  |
| CoISTA [45]| 36.58     | 0.9914   | 45.91    | 0.9963      | 39.62        | 0.9937        | 33.99  | 0.9907  |
| LMCSC-Net  | 40.31   | 0.9965   | 48.79    | 0.9981      | 41.48        | 0.9962        | 34.65  | 0.9939  |
| LMCSC +=Net | 40.81   | 0.9977   | 49.28    | 0.9989      | 43.81        | 0.9969        | 34.83  | 0.9943  |
| LMCSC-ResNet | 40.47   | 0.9974   | 47.99    | 0.9989      | 41.60        | 0.9987        | 34.96  | 0.9946  |

Table VII
Comaprison of the inference time (sec) of our multimodal architectures and CoISTA [45] w.r.t. two input resolutions.

| input size | CoISTA [45] | LMCSC-Net | LMCSC +=Net | LMCSC-ResNet |
|------------|--------------|------------|-------------|--------------|
| 256×256   | 0.61         | 1.09       | 1.14        | 1.14         |
| 512×512   | 0.63         | 1.29       | 1.38        | 1.39         |

[6] ——, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in Asian Conference on Computer Vision (ACCV), 2014.

[7] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super-resolution as sparse representation of raw image patches,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[8] K. Jia, X. Wang, and X. Tang, “Image transformation based on learning dictionaries across image spaces,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 2, pp. 367–380, 2012.

[9] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, “Convolutional sparse coding for image super-resolution,” in IEEE International Conference on Computer Vision (ICCV), 2015.

[10] K. I. Kim and Y. Kwon, “Single-image super-resolution using sparse feature representation and natural image prior.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 6, pp. 1127–1133, 2010.

[11] J. Sun, Z. Xu, and H. Y. Shum, “Image super-resolution using gradient profile prior,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[12] D. M. Nguyen, E. Tisliqian, and N. Deligiannis, “Deep learning sparse ternary projections for compressed sensing of images,” in IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2017.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2016.

[14] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in European Conference on Computer Vision (ECCV), 2016.

[15] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” IEEE Transactions on Pattern Analysis and Machine Recognition (CVPR), 2016.

[16] J. Kim, K. Woon Lee and K. Mu Lee, “Deeply-recursice convolutional network for image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] X. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in Advances in neural information Processing systems (NIPS), 2016.

[18] J. Bruna, P. Sprechmann, and Y. LeCun, “Super-resolution with deep
convolutional sufficient statistics,” CoRR, vol. abs/1511.05666, 2015.
[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[20] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.
[21] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in European Conference on Computer Vision (ECCV), 2018.
[22] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[23] Y. Tai, J. Yang, X. Liu, and C. Xu, “MmNet: A persistent memory network for image restoration,” in IEEE International Conference on Computer Vision (ICCV), 2017.
[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[25] S. Wang, L. Zhang, Y. Liang, and Q. Pan, “Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[26] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, “Supervised coupled dictionary learning with group structures for multi-modal retrieval,” in Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.
[27] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, “Semi-supervised coupled dictionary learning for person re-identification,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[28] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, “Super-resolution person re-identification with semi-coupled low-rank dictionary learning,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[29] M. Dao, N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, “Collaborative multi-sensor classification via sparsity-based representation,” IEEE Transactions on Signal Processing, vol. 64, no. 9, pp. 2400–2415, 2016.
[30] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, “Multitask driven dictionary learning for image classification,” IEEE Transactions on Image Processing, vol. 25, no. 1, pp. 24–38, 2015.
[31] N. Deligiannakis, J. F. Mota, B. Cornelis, M. R. Rodrigues, and I. Daubechies, “Multi-modal dictionary learning for image separation with application in art investigation,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 751–764, 2016.
[32] ——, “X-ray image separation via coupled dictionary learning,” in IEEE International Conference on Image Processing (ICIP), 2016.
[33] P. Song, X. Deng, J. F. Mota, N. Deligiannis, P.-L. Dragotti, and M. Rodrigues, “Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries,” IEEE Transactions on Computational Imaging, 2019.
[34] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in IEEE International Conference on Machine Learning (ICML), 2011.
[35] Y. Li, J. B. Huang, N. Ahuja, and M. H. Yang, “Deep joint image filtering,” in European Conference on Computer Vision (ECCV), 2016.
[36] H. Wu, S. Zheng, J. Zhang, and K. Huang, “Fast end-to-end trainable guided filter,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[37] K. Gregor and Y. LeCun, “Learning fast approximations of sparse coding,” in IEEE International Conference on Machine Learning (ICML), 2010.
[38] J. R. Hershey, J. L. Roux, and F. Weninger, “Deep unfolding: Model-based inspiration of novel deep architectures,” arXiv preprint arXiv:1409.2574, 2014.
[39] B. Xin, Y. Wang, W. Gao, D. Wipf, and B. Wang, “Maximal sparsity with deep networks?” in Advances in Neural Information Processing Systems (NIPS), 2016.
[40] M. Borgerding, P. Schniter, and S. Rangan, “AMP-inspired deep networks for sparse linear inverse problems,” IEEE Transactions on Signal Processing, vol. 65, no. 16, pp. 4293–4308, 2017.
[41] Y. Yang, J. Sun, H. Li, and Z. Xu, “Deep ADMM-Net for compressive sensing MRI,” in Advances in Neural Information Processing Systems (NIPS), 2016.
[42] H. Sreret and R. Giryes, “Learned convolutional sparse coding,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.
[43] J. Zhang and B. Ghanem, “ISTA-Net: Interpretable optimization-inspired

Fig. 8. Super-resolution of the test multi-spectral image “chart” with upscaling factor ×4. (a) ground-truth, (b) CSCN [44], (c) ACSC [42], (d) DJF [35] and (e) LMCSC-ResNet.

Fig. 9. Super-resolution of the test multi-spectral image “cloth” with upscaling factor ×4. (a) ground-truth, (b) CSCN [44], (c) ACSC [42], (d) DJF [35] and (e) LMCSC-ResNet.
D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, “Robust single image super-resolution via deep networks with sparse prior,” *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, 2016.

X. Deng and P. L. Dragotti, “Deep coupled ISTA network for multi-modal image super-resolution,” *IEEE Transactions on Image Processing*, vol. 29, pp. 1683–1698, 2020.

——, “Deep convolutional neural network for multi-modal image restoration and fusion,” *arXiv preprint arXiv:1910.04066*, 2019.

E. Tsiligianni and N. Deligiannis, “Deep coupled-representation learning for sparse linear inverse problems with side information,” *IEEE Signal Processing Letters*, 2019.

I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, “Multimodal image super-resolution via deep unfolding with side information,” in *European Signal Processing Conference (EUSIPCO)*, 2019.

——, “Learned multimodal convolutional sparse coding for guided image super-resolution,” in *IEEE International Conference on Image Processing (ICIP)*, 2019.

A. Sanchez-Beato and G. Pujares, “Noniterative interpolation-based super-resolution minimizing aliasing in the reconstructed image,” *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1817–1826, 2008.

F. Zhou, W. Yang, and Q. Liao, “Interpolation-based image super-resolution using multisurface fitting,” *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3312–3318, 2012.

F. Ling, Y. Du, X. Li, W. Li, F. Xiao, and Y. Zhang, “Interpolation-based super-resolution land cover mapping,” *Remote Sensing Letters*, vol. 4, no. 7, pp. 629–638, 2013.

J. Yang, Z. Lin, and S. Cohen, “Fast image super-resolution based on in-place example regression,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

W. Dong, L. Zhang, G. Shi, and X. Li, “Nonlocally centralized sparse representation for image restoration,” *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, 2012.

S. Mallat and G. Yu, “Super-resolution with sparse mixing estimators,” *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2889–2900, 2010.

Y. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, “Feedback network for image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, “Image super-resolution via dual-state recurrent networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

I. Daubechies, M. Defrise, and C. D. Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constrain,” *Communications on Pure and Applied Mathematics*, vol. 57, 11 2004.

J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2007.

K. He, J. Sun, and X. Tang, “Guided image filtering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

B. Ham, M. Cho, and J. Ponce, “Robust guided image filtering using nonconvex potentials,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 192–207, 2018.

X. Shen, Q. Yan, L. Xu, L. Ma, and J. Jia, “Multispectral joint image restoration via optimizing a scale map,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2518–2530, 2015.

S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, “Learning dynamic guidance for depth image enhancement,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

J. Tropp and S. Wright, “Computational methods for sparse solution of linear inverse problems,” *Proceedings of the IEEE*, 2010.

M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

B. Wohlberg, “Efficient convolutional sparse coding,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

F. Heide, W. Heidrich, and G. Wetzstein, “Fast and flexible convolutional sparse coding,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

J. F. C. Mota, N. Deligiannis, and M. R. D. Rodrigues, “Compressed sensing with prior information: Strategies, geometry, and bounds,” *IEEE Transactions on Information Theory*, vol. 63, pp. 4472–4496, 2017.