Designing CNNs for Multimodal Image Restoration and Fusion via Unfolding the Method of Multipliers

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

***Abstract*—Multimodal, alias, guided, image restoration is the reconstruction of a degraded image from a target modality with the aid of a high quality image from another modality. A similar task is image fusion; it refers to merging images from different modalities into a composite image. Traditional approaches for multimodal image restoration and fusion include analytical methods that are computationally expensive at inference time. Recently developed deep learning methods have shown a great performance at a reduced computational cost; however, since these methods do not incorporate prior knowledge about the problem at hand, they result in a "black box" model, that is, one can hardly say what the model has learned. In this paper, we formulate multimodal image restoration and fusion as a coupled convolutional sparse coding problem, and adopt the Method of Multipliers (MM) for its solution. Then, we use the MM-based solution to design a convolutional neural network (CNN) encoder that follows the principle of deep unfolding. To address multimodal image restoration and fusion, we design two multimodal models which employ the proposed encoder followed by an appropriately designed decoder that maps the learned representations to the desired output. Unlike most existing deep learning designs comprising multiple encoding branches followed by a concatenation or a linear combination fusion block, the proposed design provides an efficient and structured way to fuse information at different stages of the network, providing representations that can lead to accurate image reconstruction. The proposed models are applied to three image restoration tasks, as well as two image fusion tasks. Quantitative and qualitative comparisons against various state-of-the-art analytical and deep learning methods corroborate the superior performance of the proposed framework.**

***Index Terms*—Method of multipliers, deep unfolding, multimodal image restoration, image fusion, multimodal CNN.**

I. Introduction

IMAGES can be degraded by various reasons such as poor illumination conditions, noise, blurring and low resolution. Degraded images seriously affect the subjective visual effect on human eyes and may limit the performance of machine vision systems. Multimodal image processing aims at exploiting the complementary information from different image modalities to improve the quality of the given degraded images. Multimodal imaging applications can be roughly classified into two main categories, namely, multimodal image restoration and image fusion. Multimodal image restoration is concerned with the estimation of uncorrupted images from low-resolution, noisy or blurred ones with the aid of a high quality image from another modality [1], [2]. Image fusion aims at combining multiple images—possibly from different modalities—into a single one such that the fused image is more informative than any of the individual input images [3].

A careful modeling of the underlying correlation and the dependencies between the available image modalities plays a key role in multimodal signal processing. Among existing modeling approaches, sparse coding has proven effective to represent the salient information of an image, and has been widely applied to various multimodal imaging tasks [3]–[11]. Joint-sparsity-based methods rely on the assumption that different image modalities capturing the same scene may have similarities in the representation domain. Typically, multimodal sparse coding involves learned dictionaries and requires the solution of complex optimization problems both during training and inference.

Over the last decade, deep learning (DL) models have significantly outperformed analytical image restoration and fusion methods, since they can provide powerful representations of complex structures and are computationally efficient at inference [12]. For example, numerous DL designs have been proposed for single image super-resolution (SR) such as convolutional neural networks (CNNs) [13]–[15] residual architectures [16] or attention-based models [17]. Finding a mapping from a low-resolution to a high resolution image is an ill-posed problem. The design of a neural network solution relies on specific modelling assumptions — often borrowed from conventional approaches — in order to obtain a model that can learn a good mapping. For example, the work in [17] takes into account the different kind of information in images (low-frequency and high-frequency) and employs an attention mechanism to capture it. In a similar spirit, the authors of [13] assert that their CNN for image SR can be viewed as a sparse-coding based SR method with a different non-linear mapping. Nevertheless, these kind of explanations cannot abolish the "black box" nature of deep neural networks (DNNs). Most DNN models do not allow domain knowledge about the problem at hand to be incorporated into the network structure, and their theoretical foundation is underdeveloped.

Deep unfolding [18]–[20] introduced the idea of integrating domain knowledge in the form of signal priors, e.g., sparsity, into the neural network architecture. Existing designs consist of layers performing operations similar to iterative algorithms

The authors are with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), 1050 Brussels and with imec, Kapeldreef 75, 3001 Leuven, Belgium. (e-mail: {imarivan, etsiligi, bcorneli, ndeligia}@etrovub.be)

 https://doi.org/10.1109/TCSVT.2022.3163649

for sparse coding [18], [19]. For example, the Iterative Soft Thresholding Algorithm (ISTA) was unfolded into a neural network coined LISTA [18], and the ADMM-Net [19] is a single-modal deep unfolding design that relies on the Alternating Direction Method of Multipliers (ADMM). Recently, multimodal deep unfolding for image super-resolution (SR) has been presented in [21]–[25]. The network proposed in [21] employs two LISTA branches, each one computing sparse codes of the input image modalities. The representation layers of each modality are independent and fusion is performed only at a final layer by linearly combining the respective sparse codes. A different modelling approach was adopted in our previous work [22]–[24], which is based on the assumption that correlated images can have sparse representations that are similar by means of the ℓ_1-norm (coupled sparse representations). Following this assumption, the SR models presented in [22]–[24] rely on LeSITA [20], a deep unfolding design that computes sparse codes of a target modality given the sparse codes of a guidance modality. LeSITA integrates the ℓ_1-norm similarity assumption into the network architecture, performing fusion of information at every neural network layer.

Similarly to our previous work [22]–[24], in this paper, we address multimodal image restoration and fusion with coupled sparse representations, and propose a novel deep unfolding design that relies on the method of multipliers (MM). Specifically, we first formulate coupled sparse approximation as a constrained optimization problem. Then, we leverage an alternating minimization method to obtain two sub-problems; each subproblem finds the sparse representation of one modality given the sparse representation of the other modality, and vice-versa. Finally, the algorithm is translated into a multimodal convolutional neural network. Each stage of the network corresponds to a single iteration of the MM method, alternating between the computation of the sparse codes of the target and the guidance image modalities. Different from LeSITA, which was employed in our previous work [22]–[24], the proposed coupled sparse coding design performs fusion of information at every stage by computing a new intermediate representation of each modality, using a representation of the other modality computed at the same or at a previous stage. On the contrary, all the intermediate layers of LeSITA use the same representation of the guidance modality, which is provided by a separate LISTA branch. In other words, in [22]–[24], information flows from the guidance modality to the target modality and the quality of the fused representation is limited by the output of a separate branch generating the representation of the guidance modality. This limitation is addressed in this paper with an architecture that allows a symmetric bi-directional flow of information between the involved modalities.

We employ the proposed architecture for image restoration by deploying two reconstruction strategies, leading to (i) an image restoration and (ii) an image fusion multimodal CNN design. Both architectures accept as input two source images of different modalities, compute coupled sparse representations of both modalities, and perform reconstruction at a final layer. The first model for multimodal image restoration is employed to super-resolve multi-spectral as well as NIR images with the aid of high resolution RGB images, and denoise non-flash images with the aid of their flash versions. The second proposed CNN is deployed to address the tasks of multi-focus image fusion and multi-exposure image fusion. Experimental results demonstrate the superior performance of our models compared to the state of the art, showing that the proposed approach can better capture the relationship between the involved modalities by letting both modalities guide each other.

Preliminary results of this work have been presented in [26]. In this paper, we deliver the following contributions: (i) In addition to the multimodal image restoration CNN presented in [26], we introduce a second design based on the proposed coupled sparse coding network for image fusion. (ii) We apply our models to three restoration and two fusion tasks, while [26] has only addressed multimodal image SR. (iii) We perform an ablation study in order to investigate the effectiveness of the presented model and its elements.

The paper is organized as follows. Section II reviews related work, Section III presents the background for this work and Section IV presents the proposed approach. Experimental results are presented in Section VI, and conclusions are drawn in Section VII.

A word about notation: throughout the paper, vectors are denoted by boldface lower case letters and scalars by lower case letters. We utilize boldface upper case letters to denote 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

There is a vast literature of multimodal image restoration and fusion approaches and a comprehensive review of existing works is out of the scope of this paper. Next, we present the main directions and the works that are more related to ours.

A. Multimodal Image Restoration

Image restoration includes tasks such as image super-resolution (SR) and denoising. Multimodal image SR refers to the reconstruction of a high-resolution (HR) image from its low-resolution (LR) version given an HR image from another guidance modality. The purpose of multimodal denoising is to reconstruct the original image from its noisy observation given a noiseless image from another modality. Existing image restoration approaches can be divided into spatial or joint filtering methods [1], [2], [27]–[31] and transform domain methods [4]–[8], [32].

Joint filtering methods aim at transferring the salient structural information from the guidance to the target modality. Several joint filtering techniques, known as static guidance filtering, have been proposed in [1], [29], [30]. In this category, the output structure is defined by referring to the guidance image. Statistical correlations between the target and the guidance images are not taken into account. Therefore, these techniques may introduce incorrect content to the target image when there exists structural inconsistency between the modalities. On the other hand, dynamically guided methods

such as [2], [30], [31] are more robust to structural discrepancies between the input and the guidance images. For instance, the method presented in [2] includes an explicit mapping that captures the structural dependency between the input modalities. While these model-based and joint filtering methods provide strong modeling tools, they are limited in capturing the complex correlation between image modalities.

Transform domain methods have employed conventional transforms [32] and joint sparse representations [4]–[8], [33], [34]. These techniques aim at finding a mapping between the involved modalities in the representation domain. For example in [8], the authors proposed a model that considers both similarities and disparities between different modalities under the sparse representation invariance assumption. The model relies on multimodal dictionary learning and was employed for multimodal image SR [8], [35] as well as multimodal image denoising [36]. Other sparse-representation-based methods proposed to learn complex relationships between different modalities from data include [37], [38]. In [37], the authors proposed a weighted analysis sparse representation model to learn the correlation between depth and RGB images. The study in [38] presents a method for RGB-guided hyperspectral image upsampling consisting of two stages. First, a spatial upsampling stage increases the resolution of the hyperspectral image guided by the RGB image. Second, a spectrum enhancement is performed via dictionary learning resulting in a refined HR hyperspectral image.

Purely data-driven solutions for multimodal image SR are provided by multimodal deep learning approaches. Examples include the CNN based joint image filter presented in [39], the work of [40], which is a deep learning reformulation of the widely used guided image filter [29], and a multimodal dual state recurrent network with convolutional sparse priors presented in [41].

Deep unfolding networks for multimodal image SR have been proposed in [21], [25]. Both models follow an encoder-decoder structure and employ multiple LISTA [18] or convolutional LISTA [42] branches which learn sparse representations of the input modalities. In [21], a final decoding block provides the estimation of the target HR image by combining the learned representations of the source images. The authors of [25] assume that correlated images have common and unique sparse coefficients. Three convolutional LISTA branches are used to encode this information; one for the common and two for the unique information. The encoding of the common information is computed after a concatenation of low-frequency (smooth) inputs. The HR estimation is obtained by combining the common and unique feature maps at the decoder.

The deep learning design proposed in this paper adopts the assumption that the input imaging modalities have representations with many common and a few disparate coefficients, that is, they are similar by means of the ℓ_1-norm. A similar approach is followed in [35], where the authors assume that (i) an image can be split into a low-frequency and a high-frequency component, and (ii) the low-frequency components of correlated images can be represented by the same sparse codes. The same assumptions are adopted in [25]. While in these studies the low-frequency and high-frequency representations are treated separately, our model finds a single representation for each modality and imposes coupling of the multimodal representations under the ℓ_1-norm similarity constraint. Contrary to [21], [25], the proposed coupled sparse modelling approach results in a neural network model that applies fusion of information from both modalities at every encoding stage, besides the final decoding step.

B. Image Fusion

Image fusion refers to the construction of a more comprehensive single image containing complementary information from different input images. The input images are either captured by different sensors, e.g., thermal, RGB, multispectral, infrared, or by the same sensor with different parameters, such as multi-focus and multi-exposure images.

Generally, image fusion methods can be categorized into spatial and transform domain approaches. Spatial domain approaches perform image fusion by weighted averaging the source images [43]–[47]. For example, the authors of [43] compute the weight map using contrast, color saturation and well-exposedness. The Gaussian pyramid of this weight map is multiplied with the Laplacian pyramid of the multi-exposure images in order to provide the fused image. Best-exposed blocks featuring high entropy are selected to generate the fused image in [44]. Other studies [45], [46] consider post-processing of the initial pixel weights to provide a better spatial consistency among image pixels. Transform domain methods rely on effective image representations, and have employed discrete cosine transform [48], wavelet transform [49]–[51] and sparse representations [9]–[11]. These methods usually comprise three main steps, namely, decomposition, fusion and reconstruction. Typically, fusion of information is perfomed in the representation domain, therefore, the decomposition method as well as the fusion rules are important issues when following such an approach [3], [11].

In order to address the limited representation ability of conventional transforms and the computational complexity of overcomplete sparse representations, more recent efforts aim at exploiting the great representation power of DNNs. Existing works employ autoencoders [52], generative adversarial networks [53], and CNNs [54]–[63]. Similarly to transform domain methods, most DNN-based methods include a representation, a fusion and a reconstruction step. Some of these methods are applied to image patches [52], thus, they often suffer from boundary artifacts, while others are applied to whole images [60]. Besides end-to-end designs [60], which integrate the three aforementioned steps and directly produce the enhanced image, DNNs have been also used for feature extraction; the obtained features are employed in further post-processing steps [54], [58]. For instance, the network proposed in the multi-focus method presented in [58] performs focus detection which is used as a decision map; the fused image is obtained by using the decision map and the source images. An example of an end-to-end design is the multi-level convolutional neural network (MLCNN) presented in [60]. The authors of [60] assume that each image has

low-frequency and high-frequency components and propose a hierarchical CNN architecture. The network includes two CNN branches, one for each input modality, and each CNN layer corresponds to a different feature level. Similar to our design, fusion of information is performed at intermediate layers, i.e., at each feature level, besides the final reconstruction stage. Contrast to our model, each level corresponds to different information content. The exchange of information between the input modalities is a core idea adopted in [62] as well. However, in this two-branch design, each branch performs encoding of the similarities between the input modalities at the spatial domain, while our approach encodes the similarities in the representation domain. Nevertheless, the main drawback of existing DNN models is their "black box" nature.

In this paper, we rely on results from sparse representations and use deep unfolding to build an interpretable neural network architecture that can jointly learn convolutional sparse features of the source images. Reconstruction and fusion are applied in subsequent blocks, resulting in a model that can be trained end-to-end. To the best of our knowledge, this is the first deep unfolding design applied to image fusion.

III. Background

A. Sparse Coding

The problem of representing a signal $\mathbf{y} \in \mathbb{R}^n$ using only a few atoms from a dictionary $\mathbf{D} \in \mathbb{R}^{n \times M}$, $n \leq M$, is referred to as sparse coding (SC) [64]. The sparse code $\boldsymbol{\alpha} \in \mathbb{R}^M$ can be computed as the solution of the minimization problem

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \kappa\|\boldsymbol{\alpha}\|_1, \tag{1}$$

where κ is a regularization parameter, and $\|\cdot\|_1$ denotes the ℓ_1-norm which promotes sparsity. Convolutional sparse coding (CSC) is introduced as a variant of SC and is proved to be very effective for two dimensional data, e.g., images [65]. CSC is formulated as follows:

$$\min_{\mathcal{A}} \frac{1}{2}\|\mathbf{Y} - \sum_{i=1}^{m} \mathbf{D}_i * \mathbf{A}_i\|_F^2 + \kappa \sum_{i=1}^{m} \|\mathbf{A}_i\|_1, \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ represents the input image, $\mathbf{D}_i \in \mathbb{R}^{p_1 \times p_2}$, $i = 1, ..., m$, are the atoms of a convolutional dictionary $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{p_1 \times p_2 \times m}$, and $\mathbf{A}_i \in \mathbb{R}^{n_1 \times n_2}$, $i = 1, ..., m$, are the sparse feature maps w.r.t. $\boldsymbol{\mathcal{D}}$; $\|\cdot\|_F$ denotes the Frobenius norm. The ℓ_1-norm calculates the sum of absolute values of the elements in $\mathbf{A}_i$ (as if $\mathbf{A}_i$ was vectorized).

B. Sparse Coding with Side Information

According to recent studies [66], [67], correlated signals can have similar sparse representations, that is, representations with several common and a few disparate coefficients. This type of similarity in the representation domain can be mathematically expressed by the ℓ_1-norm. Let us assume that a signal $\boldsymbol{z} \in \mathbb{R}^n$ correlated with $\mathbf{y}$ is available. Assume that the so-called side information signal $\boldsymbol{z}$ has a sparse representation $\boldsymbol{\beta}$ with respect to a dictionary $\tilde{\boldsymbol{D}} \in \mathbb{R}^{n \times M}$, $n \leq M$, i.e., $\boldsymbol{z} = \tilde{\boldsymbol{D}}\boldsymbol{\beta}$. Then a sparse representation of $\mathbf{y}$ similar to $\boldsymbol{\beta}$ can be obtained via the solution of the ℓ_1-ℓ_1 minimization problem

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \kappa_1\|\boldsymbol{\alpha}\|_1 + \kappa_2\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1. \tag{3}$$

In case of convolutional sparse coding, (3) takes the form

$$\min_{\mathcal{A}} \frac{1}{2}\|\mathbf{Y} - \sum_{i=1}^{m} \mathbf{D}_i * \mathbf{A}_i\|_F^2 + \kappa_1 \sum_{i=1}^{m} \|\mathbf{A}_i\|_1, + \kappa_2 \sum_{i=1}^{m} \|\mathbf{A}_i - \mathbf{B}_i\|_1. \tag{4}$$

Both of the above problems aim to find the representation of a target signal given the representation of a guidance signal. The ℓ_1-ℓ_1 minimization approach was first applied for the reconstruction of highly correlated signals such as sequential signals which can have similar sparse representations under the same dictionary [67]. Finding efficient coupled representations of correlated signals coming from different modalities involves a coupled dictionary learning step which is computationally expensive [8].

C. Method of Multipliers

The method of multipliers (MM) [68] is an efficient algorithm for the solution of constrained optimization problems of the form

$$\min_{\mathbf{p}} f(\mathbf{p}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{p} = \mathbf{c}, \tag{5}$$

with $\mathbf{p} \in \mathbb{R}^n$ the optimization variable, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a transformation matrix and $\mathbf{c} \in \mathbb{R}^m$ a constrained parameter. The MM algorithm solves the constrained problem in (5) by minimizing the augmented Lagrangian function. Let us define the augmented Lagrangian function as

$$L(\mathbf{p}, \boldsymbol{\rho}) = f(\mathbf{p}) + \langle \boldsymbol{\rho}, \mathbf{A}\mathbf{p} - \mathbf{c} \rangle + \frac{\eta}{2}\|\mathbf{A}\mathbf{p} - \mathbf{c}\|_2^2, \tag{6}$$

with $\boldsymbol{\rho}$ the Lagrange multiplier parameter, and $\langle \cdot, \cdot \rangle$ denoting the inner product of two vectors. Each MM iteration involves the following updates:

$$\begin{cases} \mathbf{p}^{k+1} = \arg\min_{\mathbf{p}} L(\mathbf{p}, \boldsymbol{\rho}^k), \\ \boldsymbol{\rho}^{k+1} = \boldsymbol{\rho}^k + \eta(\mathbf{A}\mathbf{p}^{k+1} - \mathbf{c}). \end{cases} \tag{7}$$

Depending on the problem at hand, various methods can be used for the solution of the minimization problem in (7). Next, we discuss proximal methods.

D. Proximal Methods

Proximal methods [69] have been proposed for the solution of optimization problems of the form

$$\min_{\mathbf{p}} h(\mathbf{p}) + \lambda g(\mathbf{p}), \tag{8}$$

where $h(\cdot)$ is a differentiable convex function and $g(\cdot)$ is convex, possibly non-smooth. A proximal method iterates over

$$\mathbf{p}^{t+1} = \mathrm{Prox}_{\mu}(\mathbf{p}^t - \frac{1}{L}\nabla h(\mathbf{p}^t)), \tag{9}$$

where $L > 0$ is an upper bound on the Lipschitz constant of ∇h, and $\mathrm{Prox}_{\mu}(\cdot)$ is the proximal operator with parameter $\mu = \frac{\lambda}{L}$, defined as

$$\mathrm{Prox}_{\mu}(\mathbf{u}) = \arg\min_{\mathbf{v}} \{\frac{1}{2}\|\mathbf{v} - \mathbf{u}\|_2^2 + \mu g(\mathbf{v})\}. \tag{10}$$

IV. Coupled CSC via the Method of Multipliers

In this paper, we employ a joint sparse representation model to address multimodal image reconstruction tasks. We consider a scene captured by multiple image modalities, and assume that the acquired images are correlated. Our goal is to combine complementary information from different modalities to obtain high quality images.

Our basic modelling assumption is that correlated images can have sparse representations that are similar by means of the ℓ_1-norm. This means that the considered images have many common and a few disparate sparse coefficients. The assumption has been used in several studies in signal processing including sequential signals [67], [70], signal acquisition in sensor networks [71], or images captured by multiview cameras [72]. Specifically, we assume that the images, $\mathbf{Y}$, $\mathbf{Z}$, are decomposed as $\mathbf{Y} = \sum_{i=1}^{m} \mathbf{D}_i^Y * \mathbf{A}_i$, $\mathbf{Z} = \sum_{i=1}^{m} \mathbf{D}_i^Z * \mathbf{B}_i$, with $\mathbf{A}_i - \mathbf{B}_i$, $i = 1 \dots m$, being a sparse signal.

Given the convolutional dictionaries, $\boldsymbol{\mathcal{D}_Y}$, $\boldsymbol{\mathcal{D}_Z}$, finding the coupled convolutional sparse codes $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$ reduces to solving a constrained optimization problem of the form:

$$\begin{aligned} \min_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}}} & \sum_{i=1}^{m} \|\mathbf{A}_i\|_1 + \sum_{i=1}^{m} \|\mathbf{B}_i\|_1 + \kappa \sum_{i=1}^{m} \|\mathbf{A}_i - \mathbf{B}_i\|_1 \\ \text{s.t.} \quad & \sum_{i=1}^{m} \mathbf{D}_i^Y * \mathbf{A}_i = \mathbf{Y}, \quad \sum_{i=1}^{m} \mathbf{D}_i^Z * \mathbf{B}_i = \mathbf{Z}, \end{aligned} \tag{11}$$

where $\kappa > 0$ is a weighting parameter. The first two terms of the objective in (11) promote sparsity of the representation coefficients, while the third term expresses the similarity between images in the representation domain.

Before proceeding to the solution of (11), we would like to note that sparse reconstruction of correlated signals has been theoretically studied in [66]. This study considers that similarity between correlated modalities can be also expressed by means of the ℓ_2-norm. However, the authors have shown that for correlated signals the ℓ_1-ℓ_1 minimization problem leads to higher accuracy than the ℓ_1-ℓ_2 minimization.

We can solve (11) by taking into account the linear properties of the convolution. We replace the convolutional dictionaries with Toeplitz matrices, and the image matrices $\boldsymbol{Y}$, $\boldsymbol{Z}$ with vectorized images $\boldsymbol{y}$, $\boldsymbol{z}$. Then, (11) takes the form:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} & \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \boldsymbol{\Phi}_y\boldsymbol{\alpha} = \mathbf{y}, \quad \boldsymbol{\Phi}_z\boldsymbol{\beta} = \mathbf{z}, \end{aligned} \tag{12}$$

where $\boldsymbol{\Phi}_y$ ($\boldsymbol{\Phi}_z$) is a concatenation of Toeplitz matrices that unroll the atoms $\boldsymbol{D}_i^Y$ ($\boldsymbol{D}_i^Z$) of the convolutional dictionary $\boldsymbol{\mathcal{D}}_Y$ ($\boldsymbol{\mathcal{D}}_Z$), and $\boldsymbol{\alpha}$ ($\boldsymbol{\beta}$) is a vector containing the sparse codes of $\mathbf{y}$ ($\mathbf{z}$).

The objective in (12) is convex w.r.t. one unknown while the other is kept fixed. Therefore, we can solve (12) by alternating between the solution of the following two sub-problems:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \boldsymbol{\Phi}_y\boldsymbol{\alpha} = \mathbf{y}, \tag{13}$$

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \boldsymbol{\Phi}_z\boldsymbol{\beta} = \mathbf{z}. \tag{14}$$

Problems (13), (14) are of the form (5) and can be solved with the method of multipliers. The augmented Lagrangian function for (13) is defined as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\rho}_1) = \|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 + \boldsymbol{\rho}_1^\top(\boldsymbol{\Phi}_y\boldsymbol{\alpha} - \mathbf{y}) \\ + \frac{\eta_1}{2}\|\boldsymbol{\Phi}_y\boldsymbol{\alpha} - \mathbf{y}\|_2^2, \end{aligned} \tag{15}$$

where $\boldsymbol{\rho}_1$ is the Lagrange multiplier and η_1 a regularization parameter. According to (7), an MM iteration includes the following updates:

$$\begin{aligned} \boldsymbol{\alpha}^{k+1} = \arg\min_{\boldsymbol{\alpha}} \{ & \|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}^k\|_1 \\ & + (\boldsymbol{\rho}_1^k)^\top(\boldsymbol{\Phi}_y\boldsymbol{\alpha} - \mathbf{y}) + \frac{\eta_1}{2}\|\boldsymbol{\Phi}_y\boldsymbol{\alpha} - \mathbf{y}\|_2^2\}, \end{aligned} \tag{16}$$

$$\boldsymbol{\rho}_1^{k+1} = \boldsymbol{\rho}_1^k + \eta_1(\boldsymbol{\Phi}_y\boldsymbol{\alpha}^{k+1} - \mathbf{y}). \tag{17}$$

Similarly, for sub-problem (14), we define:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\rho}_2) = \|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 + \boldsymbol{\rho}_2^\top(\boldsymbol{\Phi}_z\boldsymbol{\beta} - \mathbf{z}) \\ + \frac{\eta_2}{2}\|\boldsymbol{\Phi}_z\boldsymbol{\beta} - \mathbf{z}\|_2^2. \end{aligned} \tag{18}$$

The updates for $\boldsymbol{\beta}$, $\boldsymbol{\rho}_2$ are given by

$$\begin{aligned} \boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}} \{ & \|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\beta}\|_1 \\ & + (\boldsymbol{\rho}_2^k)^\top(\boldsymbol{\Phi}_z\boldsymbol{\beta} - \mathbf{z}) + \frac{\eta_2}{2}\|\boldsymbol{\Phi}_z\boldsymbol{\beta} - \mathbf{z}\|_2^2\}, \end{aligned} \tag{19}$$

$$\boldsymbol{\rho}_2^{k+1} = \boldsymbol{\rho}_2^k + \eta_2(\boldsymbol{\Phi}_z\boldsymbol{\beta}^{k+1} - \mathbf{z}). \tag{20}$$

The minimization problems (16), (19) are of the form (8) and can be solved with proximal methods. For the update of $\boldsymbol{\alpha}$, we define the smooth term $h(\boldsymbol{\alpha}) = (\boldsymbol{\rho}_1^k)^\top(\boldsymbol{\Phi}_y\boldsymbol{\alpha} - \mathbf{y}) + \frac{\eta_1}{2}\|\boldsymbol{\Phi}_y\boldsymbol{\alpha} - \mathbf{y}\|_2^2$, and the non-smooth term $g(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}^k\|_1$. Then, the $(k+1)$-update of $\boldsymbol{\alpha}$ can be obtained by a proximal algorithm, computing at the $(t+1)$-th iteration:

$$\boldsymbol{\alpha}_{t+1}^{k+1} = \xi_{\mu_1}\big(\boldsymbol{\alpha}_t^{k+1} - \frac{1}{L}(\boldsymbol{\Phi}_y^\top\boldsymbol{\rho}_1^k + \eta_1\boldsymbol{\Phi}_y^\top\boldsymbol{\Phi}_y\boldsymbol{\alpha}_t^{k+1} - \eta_1\boldsymbol{\Phi}_y^\top\mathbf{y})\big). \tag{21}$$

where ξ_{μ_1} is the proximal operator, which can be obtained as the solution of (10). Nevertheless, since the proximal operator depends only on $g(\boldsymbol{\alpha})$, we can borrow ξ_{μ_1} from [20], where a problem with a similar non-smooth term was addressed; according to this study, we set $\mu_1 = \mu(1+\kappa)/2$. Note that, the analysis in [20] shows that the proximal operator expresses the correlation between the sparse signals $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$.

Repeating the same analysis for (19) results in a proximal algorithm for the $(k+1)$-update of $\boldsymbol{\beta}$. At the $(t+1)$-th iteration the algorithm computes:

$$\boldsymbol{\beta}_{t+1}^{k+1} = \xi_{\mu_2}\big(\boldsymbol{\beta}_t^{k+1} - \frac{1}{L}(\boldsymbol{\Phi}_z^\top\boldsymbol{\rho}_2^k + \eta_2\boldsymbol{\Phi}_z^\top\boldsymbol{\Phi}_z\boldsymbol{\beta}_t^{k+1} - \eta_2\boldsymbol{\Phi}_z^\top\mathbf{z})\big), \tag{22}$$

with ξ_{μ_2} parameterized by $\mu_2 = \mu(1+\kappa)/2$.

The sparse structure of the dictionaries $\boldsymbol{\Phi}_y$, $\boldsymbol{\Phi}_z$ involved in equations (17), (21), (20), (22) make the computations inefficient. We can write these equations in the form of convolutions by taking into account the Toeplitz structure of the involved dictionaries. Considering that the transpose of a Toeplitz matrix is also a Toeplitz matrix, we define $\tilde{\boldsymbol{\mathcal{D}}}_Y$ ($\tilde{\boldsymbol{\mathcal{D}}}_Z$)

Algorithm 1 MM-based coupled convolutional sparse coding

Require: Images $\mathbf{Y}$, $\mathbf{Z}$, convolutional dictionaries $\boldsymbol{\mathcal{D}}_Y$, $\boldsymbol{\mathcal{D}}_Z$
Require: # of updates $K+1$, # of iterations $T+1$
Require: Regularization parameters η_1 and η_2
Initialize $\boldsymbol{\rho}_1^0$, $\boldsymbol{\rho}_2^0$ and $\boldsymbol{\mathcal{B}}^0$
for $k = 0 : K$ **do**
  for $t = 0 : T$ **do**
    Given $\boldsymbol{\rho}_1^k$, $\boldsymbol{\mathcal{B}}^k$, $\boldsymbol{\mathcal{D}}_Y$ solve for $\boldsymbol{\mathcal{A}}^{k+1}$ using (23)
  end for
  Given $\boldsymbol{\mathcal{A}}^{k+1}$, $\boldsymbol{\mathcal{D}}_Y$ update $\boldsymbol{\rho}_1^{k_1}$ using (24)
  for $t = 0 : T$ **do**
    Given $\boldsymbol{\rho}_2^k$, $\boldsymbol{\mathcal{A}}^{k+1}$, $\boldsymbol{\mathcal{D}}_Z$ solve for $\boldsymbol{\mathcal{B}}^{k+1}$ using (25)
  end for
  Given $\boldsymbol{\mathcal{B}}^{k+1}$, $\boldsymbol{\mathcal{D}}_Z$ update $\boldsymbol{\rho}_2^{k_1}$ using (26)
end for

as the convolutional dictionary corresponding to $\boldsymbol{\Phi}_y^\top$ ($\boldsymbol{\Phi}_z^\top$). Then, (21), (17) can be written as follows:

$$\boldsymbol{\mathcal{A}}_{t+1}^{k+1} = \xi_{\mu_1}\big(\boldsymbol{\mathcal{A}}_t^{k+1} - \tilde{\boldsymbol{\mathcal{D}}}_Y * \boldsymbol{\rho}_1^k + \eta_1 \tilde{\boldsymbol{\mathcal{D}}}_Y * \boldsymbol{\mathcal{D}}_Y * \boldsymbol{\mathcal{A}}_t^{k+1} - \eta_1 \tilde{\boldsymbol{\mathcal{D}}}_Y * \mathbf{Y}\big), \tag{23}$$

$$\boldsymbol{\rho}_1^{k+1} = \boldsymbol{\rho}_1^k + \eta_1(\boldsymbol{\mathcal{D}}_Y * \boldsymbol{\mathcal{A}}_{T+1}^{k+1} - \mathbf{Y}), \tag{24}$$

with $t = 0, \ldots, T$, and $\boldsymbol{\mathcal{A}}_{T+1}^{k+1}$ the solution obtained after $T+1$ iterations of (23). Similarly, (22), (20) take the form:

$$\boldsymbol{\mathcal{B}}_{t+1}^{k+1} = \xi_{\mu_2}\big(\boldsymbol{\mathcal{B}}_t^{k+1} - \tilde{\boldsymbol{\mathcal{D}}}_Z * \boldsymbol{\rho}_2^k + \eta_2 \tilde{\boldsymbol{\mathcal{D}}}_Z * \boldsymbol{\mathcal{D}}_Z * \boldsymbol{\mathcal{B}}_t^{k+1} - \eta_2 \tilde{\boldsymbol{\mathcal{D}}}_Z * \mathbf{Z}\big). \tag{25}$$

$$\boldsymbol{\rho}_2^{k+1} = \boldsymbol{\rho}_2^k + \eta_2(\boldsymbol{\mathcal{D}}_Z * \boldsymbol{\mathcal{B}}_{T+1}^{k+1} - \mathbf{Z}). \tag{26}$$

Equations (23)–(26) describe the four steps included in a single update of an MM-based algorithm for the solution of the coupled CSC problem (11). Algorithm 1 summarizes the proposed method. Contrary to other coupled sparse coding approaches, the proposed algorithm does not rely on the sparse representation of one modality to find the sparse representation of the other modality, but alternates between the two modalities, trying to transfer similar information from one representation to the other.

We would like to note that our analysis assumes that the dictionaries $\boldsymbol{\mathcal{D}}_Y$, $\boldsymbol{\mathcal{D}}_Z$ deployed in the sparse representation of the given image modalities are predefined or known. In most joint sparse representation models, the involved dictionaries are learned from data. Therefore, an effective implementation of the proposed approach should include a dictionary learning step. Nevertheless, our goal is to employ the presented algorithm for the design of a neural network. In this case, $\boldsymbol{\mathcal{D}}_Y$, $\boldsymbol{\mathcal{D}}_Z$ are network parameters that are learned during training.

V. Multimodal CNN design

In Section IV, we presented an iterative algorithm for the computation of coupled convolutional sparse codes. Next, we propose a multimodal CNN obtained by unfolding the iterative method. We use the proposed design to obtain two multimodal neural network models tailored to specific image reconstruction tasks.

A. Unfolding MM-based Coupled CSC

In order to unfold (23) into a neural network form, we consider a single iteration ($T = 0$) of the proximal algorithm in each update of $\boldsymbol{\mathcal{A}}$, and set $\boldsymbol{\mathcal{A}}^{k+1} := \boldsymbol{\mathcal{A}}_1^{k+1}$. We also set $\eta_1 = 1$. Then, we rewrite (23) as follows:

$$\boldsymbol{\mathcal{A}}^{k+1} = \xi_{\mu_1}\big(\boldsymbol{\mathcal{A}}^k - \boldsymbol{\mathcal{Q}}_1 * \boldsymbol{\rho}_1^k + \boldsymbol{\mathcal{S}}_1 * \boldsymbol{\mathcal{A}}^k - \boldsymbol{\mathcal{R}}_1 * \mathbf{Y}\big), \tag{27}$$

with $\boldsymbol{\mathcal{Q}}_1 := \tilde{\boldsymbol{\mathcal{D}}}_Y$, $\boldsymbol{\mathcal{S}}_1 := \tilde{\boldsymbol{\mathcal{D}}}_Y * \boldsymbol{\mathcal{D}}_Y$, $\boldsymbol{\mathcal{R}}_1 := \tilde{\boldsymbol{\mathcal{D}}}_Y$. The convolutional terms $\boldsymbol{\mathcal{Q}}_1 * \boldsymbol{\rho}_1^k$, $\boldsymbol{\mathcal{S}}_1 * \boldsymbol{\mathcal{A}}^k$ and $\boldsymbol{\mathcal{R}}_1 * \mathbf{Y}$ can be implemented as convolutional neural network layers with no activation function and a stride size of 1 at each direction. Although it is possible to design convolutional filters with other stride sizes, in what follows we only consider the stride size of 1 for simplicity. Zero padding is performed at each layer in order to preserve the spatial resolution throughout the network. The three convolutional layers are followed by an activation function implementing the proximal operator $\xi_{\mu_1}(\cdot)$ [20]. For the update of $\boldsymbol{\rho}_1$, (26) can be written as follows:

$$\boldsymbol{\rho}_1^{k+1} = \boldsymbol{\rho}_1^k + \boldsymbol{\mathcal{T}}_1 * \boldsymbol{\mathcal{A}}^{k+1} - \mathbf{Y}, \tag{28}$$

with $\boldsymbol{\mathcal{T}}_1 := \boldsymbol{\mathcal{D}}_Y$. Equation (28) can be also unrolled into a neural network form. Figure 1 depicts the updates of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\rho}_1$ implemented by a CNN block.

Concerning the updates of $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\rho}_2$, we rewrite (25), (26) as follows:

$$\boldsymbol{\mathcal{B}}^{k+1} = \xi_{\mu_2}\big(\boldsymbol{\mathcal{B}}^k - \boldsymbol{\mathcal{Q}}_2 * \boldsymbol{\rho}_2^k + \boldsymbol{\mathcal{S}}_2 * \boldsymbol{\mathcal{B}}^k - \boldsymbol{\mathcal{R}}_2 * \mathbf{Z}\big), \tag{29}$$

$$\boldsymbol{\rho}_2^{k+1} = \boldsymbol{\rho}_2^k + \boldsymbol{\mathcal{T}}_2 * \boldsymbol{\mathcal{B}}^{k+1} - \mathbf{Z}. \tag{30}$$

Unrolling these equations into a neural network form results in a structure similar to the one presented in Fig. 1 (we replace $\mathbf{Y}$ with $\mathbf{Z}$, and $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\rho}_1$ with $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\rho}_2$, respectively).

In what follows, a neural network block computing a single update of the coupled convolutional sparse codes $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$ is referred to as a coupled CSC (C-CSC) stage. A C-CSC stage is parameterized by $\boldsymbol{\mathcal{Q}}_i, \boldsymbol{\mathcal{R}}_i, \boldsymbol{\mathcal{S}}_i, \boldsymbol{\mathcal{T}}_i$, $i = 1, 2$; the parameters can be learned from data. Repeating several C-CSC stages yields a multimodal deep neural network performing a fixed number of iterations of the MM-based coupled-CSC algorithm. Learning the network parameters can result in accurate estimation of the sparse codes with only a few C-CSC stages.

We would like to note that the proposed design performs coupling of the involved coefficients by (i) alternating between the updates of $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$, and (ii) by using an activation function that implements the proximal operators ξ_{μ_1}, ξ_{μ_2}, which also express the similarity between $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$. Therefore, the proposed architecture can effectively capture the correlation between the input images.

The alternating update of the sparse codes of both modalities implemented by each C-CSC stage is a key feature of our design. Even in case the guidance modality $\mathbf{Z}$ is of high quality, the bidirectional flow of information between the two modalities is necessary for successful coupling. Images $\mathbf{Y}$ and $\mathbf{Z}$ can have different sparse representations under different convolutional dictionaries. Similar to multimodal dictionary learning, finding a representation with several common coefficients requires to alternate between modalities to achieve

both efficient coupling and approximation accuracy. To put it differently, suppose we find a representation of $\mathbf{Z}$ and force $\mathbf{Y}$ to follow this representation. Then, the results might not be as good as possible either in terms of ℓ_1-norm similarity or in terms of approximation accuracy. In Section VI-D, we compare the proposed bidirectional architecture with a similar single-directional design and show the effectiveness of our approach.

The bidirectional flow of information adopted in the proposed design is a key difference with the designs presented in our previous work [22]–[24]. These models employ a separate (LISTA or convolutional LISTA) neural network branch to generate the encoding of the guidance modality and force the target modality to follow this encoding. The main advantage of the architecture proposed in this paper compared to our previous work is that both modalities are treated in the same manner. This is achieved by the mathematical formulation introduced by (11). While our previous work relies on the solution of an ℓ_1-ℓ_1 minimization problem of the form (3) (in [22]) or (4) (in [23], [24]), which find a representation of the target modality given the representation of the guidance modality, the algorithm proposed in this paper solves (11) where both representations are considered unknown. Furthermore, the algorithms unrolled in our previous work [22]–[24] are based on the proximal method, whereas the network architecture in this work is based on the method of multipliers [68], thereby resulting in a different form of layers. Even though, both approaches employ the same mechanism (activation function) to force one encoding follow the other, the alternating steps implemented by the proposed architecture together with the different form of layers offer more flexibility and result in more efficient coupling. This is corroborated by experimental results presented in Section VI.

B. Training the proposed CNN

The proposed design can learn to predict coupled sparse codes of the input imaging modalities. Let us denote as $\boldsymbol{\mathcal{C}} = f(\Theta, \boldsymbol{Y}, \boldsymbol{Z})$ the proposed multimodal encoder, where $\boldsymbol{\mathcal{C}} = [\boldsymbol{\mathcal{A}}\ \boldsymbol{\mathcal{B}}]$, and Θ is the set of all trainable parameters of the proposed architecture. Then, training with gradient-based learning methods is feasible if f is continuous and almost-everywhere differentiable with respect to Θ. f includes the non-linear activations ξ_{μ_1}, ξ_{μ_2} which are piecewise linear functions [20]. Although ξ_{μ_1}, ξ_{μ_2} are not differentiable at some points, at these points the subderivatives can be easily computed; therefore, the multimodal encoder can be trained with gradient-based methods.

We obtain a training dataset by executing several iterations of Algorithm 1 on J pairs of correlated images $\{\boldsymbol{Y}_j, \boldsymbol{Z}_j\}_{j=1}^J$. Let $\{\boldsymbol{\mathcal{A}}_j^*, \boldsymbol{\mathcal{B}}_j^*\}_{j=1}^J$ denote the set of the corresponding sparse codes. Then, training can be performed by minimizing the squared error between the available and the predicted sparse codes, that is,

$$\mathcal{L}_{\text{sparse}}(\Theta, \boldsymbol{Y}_j, \boldsymbol{Z}_j) = \frac{1}{2}\|\boldsymbol{\mathcal{A}}_j^* - \hat{\boldsymbol{\mathcal{A}}}_j\|_2^2 + \frac{1}{2}\|\boldsymbol{\mathcal{B}}_j^* - \hat{\boldsymbol{\mathcal{B}}}_j\|_2^2, \quad (31)$$

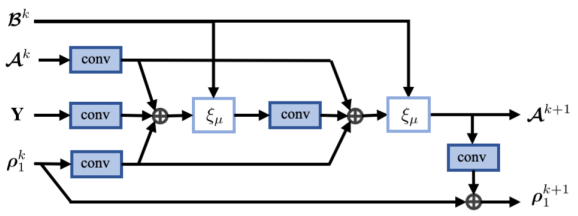


Fig. 1. A CNN block for the computation of the convolutional sparse codes $\boldsymbol{\mathcal{A}}^{k+1}$ of an input image $\mathbf{Y}$, given the convolutional sparse codes $\boldsymbol{\mathcal{B}}^k$ of a correlated image $\mathbf{Z}$; $\boldsymbol{\rho}_1^k$ is a parameter.

where $[\hat{\boldsymbol{\mathcal{A}}}_j\ \hat{\boldsymbol{\mathcal{B}}}_j] = f(\Theta, \boldsymbol{Y}_j, \boldsymbol{Z}_j)$. The mean squared error (MSE) is computed over the whole training set. By employing stochastic gradient descent we obtain the t-th updating step:

$$\Theta_{t+1} = \Theta_t - \eta \frac{\partial \mathcal{L}_{\text{sparse}}(\Theta, \boldsymbol{Y}_t, \boldsymbol{Z}_t)}{\partial \Theta}, \quad (32)$$

where η is the learning rate.

Next, we apply the proposed CNN design to address guided image restoration and image fusion.

C. Multimodal Image Restoration

In multimodal or guided image restoration, we aim at reconstructing a high quality image $\mathbf{X}$, given a distorted (noisy and/or low-resolution) image $\mathbf{Y}$ from the target modality and a clean and high quality image $\mathbf{Z}$ from a second modality; the second image modality is used to guide the restoration process. We address this problem by making the following modelling assumptions: (i) The input image $\mathbf{Y}$ and the output image $\mathbf{X}$ share the same sparse feature maps w.r.t. different convolutional dictionaries, that is, $\mathbf{X} = \sum_{i=1}^m \mathbf{D}_i^X * \mathbf{A}_i$, $\mathbf{Y} = \sum_{i=1}^m \mathbf{D}_i^Y * \mathbf{A}_i$. Therefore, the recontruction of $\mathbf{X}$ reduces to finding the sparse feature maps of $\mathbf{Y}$, assuming that the dictionaries are given. (ii) Images from different modalities capturing the same scene have similar sparse representations. Specifically, we assume that the input image $\mathbf{Y}$ decomposed as $\mathbf{Y} = \sum_{i=1}^m \mathbf{D}_i^Y * \mathbf{A}_i$, and the guidance image $\mathbf{Z}$, decomposed as $\mathbf{Z} = \sum_{i=1}^m \mathbf{D}_i^Z * \mathbf{B}_i$, have sparse feature maps that are similar by means of the ℓ_1-norm. Therefore, image restoration can be formulated as a problem of the form (11).

We propose a multimodal CNN for guided image restoration comprising several C-CSC stages that estimate the sparse codes of the input $\mathbf{Y}$ guided by the sparse codes of $\mathbf{Z}$. In the last layer, we deploy a convolutional dictionary in order to reconstruct the desired image $\mathbf{X}$. Figure 2 illustrates the proposed multimodal CNN design with three C-CSC stages. We train the network end-to-end by minimizing the MSE loss function

$$\mathcal{L} = \frac{1}{J}\sum_{j=1}^J \|\boldsymbol{X}_j - \hat{\boldsymbol{X}}_j\|_2^2, \quad (33)$$

where $\boldsymbol{X}_j$, $\hat{\boldsymbol{X}}_j$, $j = 1, \ldots, J$, are the ground-truth and estimated images, respectively, and J is the number of the available training samples.

In Section VI, we deploy the proposed CNN for guided image super-resolution and guided image denoising. For the first task, we assume that a low-resolution (LR) image $\mathbf{Y}$

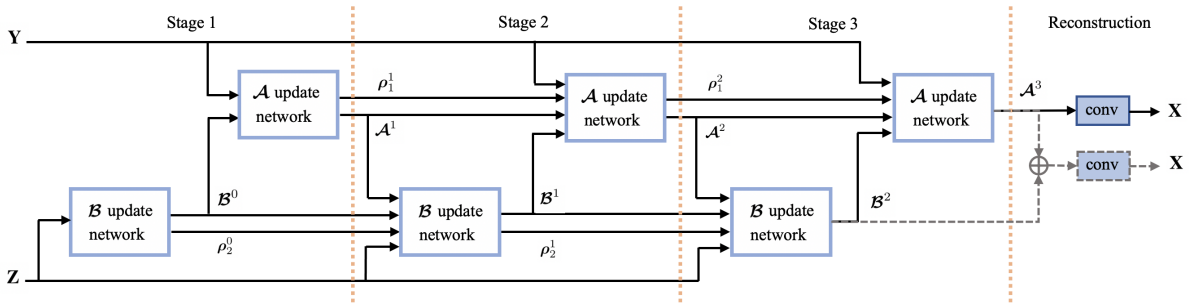


Fig. 2. The proposed multimodal MM-based CNN using three stages and one reconstruction layer. The black solid lines in the last layer illustrate the output of the restoration CNN while the dashed lines represent the fusion network. The "plus" sign in the fusion layer represents a linear combination of the features.

from a target modality, and a high-resolution (HR) image $\mathbf{Z}$ from a guidance modality have similar sparse representations by means of ℓ_1-norm; the LR image $\mathbf{Y}$ and the unknown HR image $\mathbf{X}$ of the target modality share the same sparse representation. For image denoising, we assume that the noisy and denoised images of the target modality share the same sparse feature maps.

D. Image Fusion

In image fusion, we aim at reconstructing a high quality image $\mathbf{X}$, given two degraded images $\mathbf{Y}$, $\mathbf{Z}$, possibly from different modalities. We address this problem in a similar way to guided image restoration. We adopt a joint sparse representation model and assume that two complementary source images, when represented w.r.t. appropriate dictionaries, can have similar sparse representations by means of the ℓ_1-norm. We obtain the coupled convolutional sparse codes, $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$, of the two given images $\mathbf{Y}$, $\mathbf{Z}$ as the solution of (11). Having computed $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$, the fused image can be obtained by averaging the reconstructed images from the two modalities, that is, $\mathbf{X} = \lambda_1 \sum_{i=1}^{m} \mathbf{D}_i^Y * \mathbf{A}_i + \lambda_2 \sum_{i=1}^{m} \mathbf{D}_i^Z * \mathbf{B}_i$, with $\lambda_1 + \lambda_2 = 1$. Since both input modalities are degraded, averaging allows the model to exploit the high-frequency information (details) carried by each modality.

We use this modelling assumption to design a multimodal CNN as follows. We build a network comprising several C-CSC stages followed by a reconstruction step for each modality. Then, the fused image is obtained as a sum of the reconstructed images of both modalities (we assume that the averaging weights λ_1, λ_2 are absorbed by the reconstruction dictionaries). Figure 2 illustrates the proposed architecture with three C-CSC stages. Similar to our previous design, the fusion network is trained end-to-end using the MSE loss function (33). In Section VI, we evaluate the proposed network on two fusion tasks, namely, multi-focus image fusion and multi-exposure image fusion.

VI. Experiments

This section presents the implementation details for the proposed model and its performance evaluation. First, the model is employed for image restoration, namely, the super-resolution of multi-spectral and NIR data with the aid of HR RGB images, and denoising of flash/non-flash images. Second, we apply the proposed model for multi-focus and multi-exposure image fusion. The experiments include comparison with state-of-the-art methods. The superior performance of our approach is demonstrated both by numerical and visual results.

A. Experimental Setting

We realize a model with three C-CSC stages similar to the one depicted in Fig. 2. We initialize the representation $\boldsymbol{\mathcal{A}}$ of the target modality and the Lagrange parameters $\boldsymbol{\rho_1}$, $\boldsymbol{\rho_2}$ with zero. The convolutional layers contain 16 7×7 kernels. Moreover, zero padding is applied to the input of each convolutional layer to preserve the same spatial size throughout the model. Initial values of the convolutional kernels are randomly drawn from a Gaussian distribution with a standard deviation 0.01. The parameters μ_1, μ_2 of the proximal operators are initialized to 0.2. We train the network using the Adam optimizer with learning rate 0.0001 and mini-batch size 32.

B. Multimodal Image Restoration

1) Multimodal image super-resolution: In multimodal image SR, the inputs to the network are the LR image $\mathbf{Y}$ from the target modality and an HR image $\mathbf{Z}$ from the guidance modality. The network provides the reconstructed HR target image with the aid of the HR guidance image. The LR images are obtained by performing blurring and downscaling operations on the ground truth images. The guidance modality in our network only includes the luminance channel of the corresponding RGB image. We utilize two multimodal datasets, namely, the Columbia multi-spectral database[1] and the EPFL RGB-NIR dataset[2]. We reserve seven pairs from the multispectral dataset and eight pairs form the NIR dataset for testing. We apply SR at scales of $\times 4$ and $\times 8$ for the multispectral data while choosing scales of $\times 2$ and $\times 4$ for the NIR data in order to provide a wider diversity of the results. We train the network separately for every scale and dataset.

For the experiments with multispectral data, we create a training set consisting of 64×64 image patches and apply data augmentation with rotation obtaining a dataset of size $40,000$.

[1] http://www.cs.columbia.edu/CAVE/databases/multispectral
[2] https://ivrl.epfl.ch/supplementary_material/cvpr11/

TABLE I
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 $\times 4$ AND $\times 8$ UPSCALING FACTORS.

MS/RGB	Chart toy		Egyptian		Feathers		Glass tiles		Jelly beans		Oil Paintings		Paints		Average	
$\times 4$	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	28.94	0.9424	36.57	0.9786	30.80	0.9562	26.65	0.9242	27.81	0.9302	31.67	0.8943	29.29	0.9493	30.25	0.9393
SDF [30]	31.87	0.9694	39.43	0.9795	33.45	0.9650	28.22	0.9374	30.32	0.9433	32.86	0.9126	31.96	0.9655	32.59	0.9532
JBF [1]	32.56	0.9653	38.73	0.9735	33.60	0.9637	27.52	0.9341	30.29	0.9498	32.77	0.8962	31.94	0.9699	32.49	0.9504
JFSM [2]	32.98	0.9295	40.39	0.9705	33.89	0.9425	28.98	0.9397	31.18	0.9451	35.91	0.9560	32.76	0.9430	33.73	0.9466
GF [29]	34.09	0.9788	40.24	0.9796	33.60	0.9748	29.46	0.9593	30.90	0.9658	35.03	0.9441	31.73	0.9702	33.58	0.9675
SRCNN [13]	31.29	0.9796	38.04	0.9803	33.43	0.9848	27.85	0.9689	30.78	0.9771	30.50	0.9153	34.64	0.9878	32.36	0.9705
FSRCNN [14]	30.43	0.9770	38.29	0.9862	32.72	0.98418	28.15	0.9690	29.74	0.9781	32.27	0.9606	32.28	0.9844	31.98	0.9771
EDSR [16]	33.45	0.9836	40.03	0.9829	35.55	0.9875	29.75	0.9736	32.81	0.9838	32.69	0.9178	37.28	0.9914	34.51	0.9739
SRFBN [73]	33.43	0.9838	40.04	0.9822	35.53	0.9873	29.53	0.9676	32.97	0.9845	32.68	0.9182	36.06	0.9907	34.32	0.9735
DRN [74]	33.62	0.9844	41.01	0.9868	35.75	0.9893	29.57	0.9678	33.03	0.9856	32.93	0.9187	37.24	0.9914	34.73	0.9748
DGF [40]	34.19	0.9559	37.81	0.9620	31.22	0.9336	29.93	0.9339	28.94	0.9459	36.10	0.9649	31.72	0.9680	32.84	0.9520
DJF [39]	37.86	0.9935	45.69	0.9922	40.13	0.9939	34.97	0.9915	39.16	0.9885	37.76	0.9805	39.36	0.9944	39.28	0.9906
CoISTA [21]	36.58	0.9914	45.91	0.9961	39.62	0.9937	33.99	0.9907	38.92	0.9956	37.26	0.9690	38.40	0.9949	38.67	0.9902
LMCSC [24]	40.31	0.9965	48.79	0.9981	41.48	0.9962	34.65	0.9939	39.75	0.9966	39.14	0.9910	38.98	0.9966	40.44	0.9955
CU-Net [25]	39.47	0.9960	46.48	0.9926	**42.43**	0.9964	**36.03**	0.9944	40.64	0.9972	38.90	0.9840	**40.90**	0.9971	40.69	0.9940
proposed	**41.13**	**0.9977**	**49.92**	**0.9991**	42.27	**0.9974**	34.84	**0.9948**	**40.75**	**0.9980**	**39.84**	**0.9926**	39.69	**0.9978**	**41.20**	**0.9968**
$\times 8$	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	25.00	0.8048	33.12	0.9316	25.59	0.8067	22.56	0.7308	23.04	0.7388	30.56	0.8234	26.40	0.8465	26.61	0.8118
SDF [30]	27.89	0.8933	34.22	0.9366	28.25	0.8967	24.66	0.8253	24.88	0.8527	31.61	0.8587	28.54	0.9246	28.58	0.8840
JBF [1]	28.00	0.9092	35.92	0.9578	26.48	0.8492	23.06	0.7706	24.74	0.8532	31.90	0.8531	28.21	0.9135	28.33	0.8723
JFSM [2]	30.28	0.8897	38.31	0.9451	28.24	0.9043	25.31	0.8851	26.59	0.9083	32.55	0.9199	29.97	0.9339	30.18	0.9123
GF [29]	29.77	0.9446	36.52	0.9627	27.85	0.9183	25.24	0.8641	25.05	0.8813	33.82	0.9383	29.25	0.9451	29.64	0.9221
SRCNN [13]	25.89	0.8768	34.26	0.9487	27.12	0.9008	22.67	0.8291	23.66	0.8214	31.06	0.8588	29.47	0.9354	27.73	0.8815
FSRCNN [14]	25.61	0.8804	33.03	0.9532	26.81	0.9147	23.88	0.8531	23.89	0.8732	29.67	0.8943	25.73	0.9186	26.95	0.8982
EDSR [16]	27.04	0.8802	35.62	0.9511	28.38	0.9040	23.53	0.8339	24.73	0.8259	31.95	0.8632	30.33	0.9425	28.78	0.8858
SRFBN [73]	27.90	0.8736	35.50	0.9755	30.14	0.9718	23.72	0.9002	25.80	0.9243	31.07	0.9258	28.03	0.9653	28.88	0.9338
DRN [74]	28.16	0.8748	35.72	0.9771	30.48	0.9732	24.02	0.9036	25.96	0.9262	31.33	0.9273	28.51	0.9669	29.16	0.9355
DGF [40]	28.39	0.8901	34.98	0.9444	26.83	0.8353	25.46	0.8987	25.97	0.8852	33.25	0.9351	27.89	0.9128	28.97	0.9002
DJF [39]	32.89	0.9733	41.58	0.9850	31.50	0.9396	29.53	0.9685	30.14	0.9503	35.12	0.9492	31.86	0.9553	33.23	0.9602
CoISTA [21]	33.18	0.9768	43.46	0.9906	32.04	0.9493	27.96	0.9390	30.69	0.9585	35.99	0.9482	33.05	0.9679	33.77	0.9615
LMCSC [24]	34.35	0.9805	43.90	0.9966	36.81	0.9875	30.20	0.9724	34.70	0.9888	36.27	0.9759	**35.06**	0.9910	35.90	0.9847
CU-Net [25]	34.21	0.9760	42.91	0.9952	36.27	0.9851	30.51	0.9697	33.94	0.9858	36.59	0.9749	34.75	0.9894	35.58	0.9823
proposed	**35.47**	**0.9837**	**44.40**	**0.9969**	**37.83**	**0.9901**	**30.89**	**0.9797**	**34.79**	**0.9890**	**36.98**	**0.9793**	34.96	**0.9927**	**36.47**	**0.9873**

We compare the multispectral image SR results against single-modal deep learning methods such as SRCNN [13], FSRCNN [14], EDSR [16], SRFBN [73], DRN [74], and several multimodal image SR techniques including optimization based approaches, e.g., SDF [30], JBF [1], JFSM [2], GF [29], learning based models, e.g., DGF [40], DJF [39], CoISTA [21], CU-Net [25] and our previous work LMCSC-Net [24][3]. Table I presents numerical results in terms of Peak-Signal-to-Noise-Ratio (PSNR) and structural similarity index (SSIM). Recall that the PSNR between the ground truth image X (8 bits) and the reconstructed image $\hat{X}$ is given by $\mathrm{PSNR}(X, \hat{X}) = 20 \log_{10} \big(255/\mathrm{RMSE}\big)$, where RMSE is the root mean squared error of X, $\hat{X}$. SSIM is defined as

$$\mathrm{SSIM}(X, \hat{X}) = \frac{(2\mu_X \mu_{\hat{X}} + c_1)(2\sigma_{X\hat{X}} + c_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + c_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + c_2)},$$

where μ_X ($\mu_{\hat{X}}$) and σ_X ($\sigma_{\hat{X}}$) are the mean and the variance of image X ($\hat{X}$), respectively, and $\sigma_{X\hat{X}}$ is the covariance of X and $\hat{X}$; c_1 and c_2 are constants. As can be seen in Table I, the proposed network outperforms the abovementioned methods with a PSNR gain up to 0.57 dB compared to the second best methods, i.e., CU-Net [25] for scale $\times 4$, and LMCSC-Net [24] for scale $\times 8$. Figure 3 provides a visual comparison between the proposed and the baseline methods. Besides the reconstructed images, the figure depicts the corresponding error maps which clearly show the superior performance of the proposed approach.

For the experiments with NIR data, we use a training set containing $28,000$ pairs of image patches. We compare our results against several single-modal methods, e.g., SRCNN [13], FSRCNN [14], CSCN [75], ACSC [42], EDSR [16], SRFBN [73], DRN [74], and various multimodal models including SDF [30], DJF [39], DMSC [22], LMCSC-Net [24] and CU-Net [25]. As can be seen in Table II, the PSNR gain against the second best method, that is, CU-Net [25] is up to 0.87 dB. The superior performance of the proposed model relies not only on learning an efficient coupled representation of the modalities, but on the effective fusion of the representations through the bidirectional flow of information between the two modalities as well. In LMCSC-Net [24], a limitation that affects the performance is that the guidance representation, which is obtained using a side information branch, is kept fixed during the computation of the target modality features. Figure 4 depicts a reconstruction

[3]In [24], we have presented several architectures based on LMCSC-Net. For a fair comparison, here, we present results of the baseline model. Since the experiments include the same datasets, the interested reader can refer to [24] for further comparisons.

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 $\times 2$ AND $\times 4$ UPSCALING FACTORS.

NIR/RGB	u-0004		u-0006		u-0017		o-0018		u-0020		u-0026		o-0030		u-0050		Average	
$\times 2$	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	30.55	0.9290	36.89	0.9462	34.83	0.9183	30.50	0.9116	32.68	0.9324	30.51	0.9142	30.90	0.8876	30.68	0.9229	32.19	0.9202
SDF [30]	30.72	0.9290	36.71	0.9364	34.89	0.9139	30.74	0.9128	32.89	0.9317	30.58	0.9110	31.02	0.8816	30.61	0.9125	32.28	0.9161
CSCN [75]	32.77	0.9715	39.47	0.9715	36.76	0.9574	33.98	0.9659	35.54	0.9658	32.94	0.9339	33.34	0.9465	33.31	0.9693	34.76	0.9602
ACSC [42]	33.24	0.9723	39.78	0.9718	36.64	0.9579	34.26	0.9670	35.65	0.9660	33.11	0.9416	33.32	0.9465	33.39	0.9696	34.93	0.9608
SRCNN [13]	33.41	0.9729	39.86	0.9724	36.82	0.9590	34.43	0.9676	35.77	0.9668	33.32	0.9425	33.46	0.9473	33.54	0.9704	35.07	0.9623
FSRCNN [14]	32.72	0.9949	39.05	0.9961	36.28	0.9923	33.84	0.9950	35.47	0.9953	32.91	0.9941	33.19	0.9910	33.34	0.9939	34.60	0.9941
EDSR [16]	35.10	0.9971	40.66	0.9958	37.57	0.9942	35.97	0.9961	37.11	0.9968	34.21	0.9941	34.82	0.9939	36.55	0.9956	36.49	0.9954
SRFBN [73]	35.36	0.9974	41.08	0.9970	38.19	0.9950	36.47	0.9971	37.50	0.9969	31.00	0.9782	35.57	0.9944	**37.06**	**0.9966**	36.53	0.9941
DRN [74]	35.54	0.9983	41.15	0.9970	38.12	0.9947	36.67	0.9968	37.42	0.9968	34.36	0.9949	35.34	0.9942	36.81	0.9960	36.92	0.9960
DJF [39]	34.50	0.9964	41.52	0.9975	38.65	0.9961	34.78	0.9960	37.35	0.9973	33.15	0.9939	35.67	0.9944	32.60	0.9928	36.03	0.9955
DMSC [22]	36.97	0.9976	43.22	0.9977	40.41	0.9970	37.90	0.9964	40.07	0.9975	34.96	0.9948	37.74	0.9953	33.78	0.9934	38.13	0.9962
LMCSC [24]	37.26	0.9977	43.60	0.9982	40.87	0.9967	39.21	0.9982	40.98	0.9980	35.60	0.9963	38.29	0.9961	34.11	0.9948	38.74	0.9970
CU-Net [25]	37.91	0.9971	43.51	0.9975	41.15	0.9956	38.56	0.9972	40.87	0.9973	35.94	0.9949	38.67	0.9950	34.51	0.9941	38.90	0.9961
proposed	**38.09**	**0.9982**	**43.91**	**0.9985**	**41.52**	**0.9977**	**41.76**	**0.9986**	**42.17**	**0.9987**	**36.75**	**0.9969**	**39.23**	**0.9970**	34.74	0.9959	**39.77**	**0.9976**
$\times 4$	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	25.93	0.9029	30.89	0.9458	30.45	0.9527	25.19	0.9298	28.03	0.9577	26.27	0.8704	26.54	0.8401	26.65	0.9434	27.49	0.9179
SDF [30]	26.82	0.9066	30.60	0.8918	30.72	0.9281	26.09	0.9169	29.09	0.9505	26.61	0.8558	27.21	0.8415	27.07	0.9207	28.03	0.9018
CSCN [75]	27.64	0.9378	32.60	0.9361	32.60	0.9361	27.28	0.9250	30.04	0.9487	27.91	0.8724	27.72	0.8378	28.20	0.9101	29.14	0.9111
ACSC [42]	27.28	0.9371	32.61	0.9360	31.66	0.9208	27.42	0.9252	29.87	0.9483	27.92	0.8722	27.66	0.8381	27.80	0.9086	29.03	0.9107
SRCNN [13]	27.42	0.9736	32.57	0.9661	31.29	0.9674	27.64	0.9702	29.90	0.9765	27.76	0.9671	28.31	0.9513	28.57	0.9724	29.18	0.9680
FSRCNN [14]	27.34	0.9636	32.19	0.9596	31.46	0.9585	26.98	0.95362	29.57	0.9692	27.56	0.9524	27.53	0.9346	28.04	0.9579	28.84	0.9562
EDSR [16]	28.59	0.9759	33.42	0.9684	32.50	0.9693	28.54	0.9717	31.09	0.9789	28.74	0.9678	28.81	0.9528	29.58	0.9742	30.15	0.9699
SRFBN [73]	29.01	0.9787	33.73	0.9702	32.91	0.9725	28.88	0.9740	31.44	0.9807	29.10	0.9702	29.45	0.9583	29.89	0.9762	30.55	0.9726
DRN [74]	29.12	0.9793	33.78	0.9709	33.17	0.9732	29.04	0.9753	31.51	0.9815	29.23	0.9716	29.66	0.9587	30.29	0.9776	30.72	0.9735
DJF [39]	31.02	0.9784	36.04	0.9894	34.18	0.9815	30.72	0.9888	33.60	0.9915	29.21	0.9397	31.27	0.9345	28.58	0.9616	31.83	0.9707
DMSC [22]	33.19	0.9846	37.69	0.9892	36.00	0.9812	33.84	0.9888	36.33	0.9893	30.65	0.9725	33.19	0.9727	29.85	0.9716	33.84	0.9812
LMCSC [24]	33.75	0.9869	38.74	0.9912	36.16	0.9828	34.17	0.9902	36.95	0.9900	31.03	0.9784	33.56	0.9780	30.04	0.9772	34.28	0.9843
CU-Net [25]	34.08	0.9919	38.74	0.9928	36.18	0.9858	34.96	0.9935	37.39	0.9938	31.10	0.9815	33.59	0.9816	29.82	0.9773	34.49	0.9873
proposed	**34.21**	**0.9923**	**39.77**	**0.9938**	**36.84**	**0.9873**	**36.51**	**0.9942**	**37.58**	**0.9941**	**31.76**	**0.9837**	**33.89**	**0.9822**	**30.41**	**0.9793**	**35.12**	**0.9883**

example with the corresponding error maps, providing a visual comparison between the proposed and the baseline methods.

2) Multimodal image denoising: We also apply the proposed network to the denoising of non-flash images guided by their flash counterparts. To this end, we employ the flash/non-flash dataset presented in [76]. We reserve 12 image pairs from the dataset for testing and extract patches of size 64×64 from 200 image pairs, obtaining a training set of size $50,000$. We add Gaussian noise with three different levels of $\sigma = 25, 50$ and 75 to the non-flash images. We evaluate the performance of the proposed network by comparing against two single-modal image denoising methods, i.e., CBM3D [77], DnCNN [78], and three multimodal methods, i.e., DJF [39], MuGIF [31] and CU-Net [25]. The proposed fusion strategy provides an enhanced performance as can be seen from the numerical results presented in Table III.

C. Image Fusion

1) Multi-focus image fusion: In multi-focus image fusion, we aim at producing an all-in-focus image from several images with different depth of focus. As there is no dataset consisting of near-, far- and all-in-focus images with a proper size, we synthesize a training dataset. For this purpose, we utilize the images from the General-100 dataset and generate two types of a focused image by randomly selecting the foreground/background area and blurring one of them for each image. The number of training data is $30,000$. For testing, we use the multi-focus image from the Lytro dataset [79]. Our model is compared with the DCT-Corr [48], which is a transform-domain method, and the deep learning designs presented in [58] and [63].

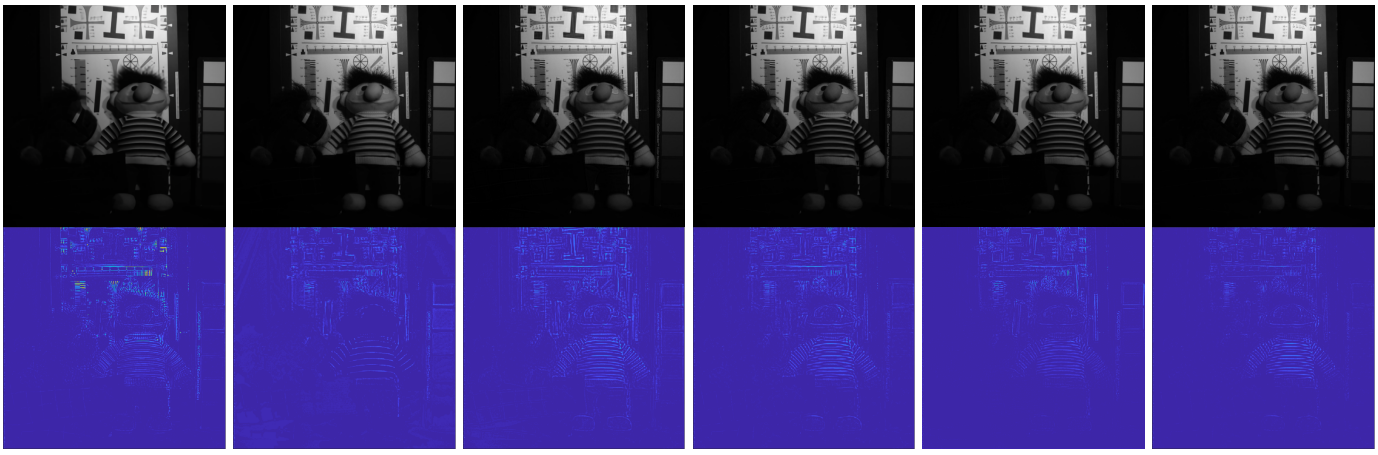
We present visual results of this fusion task in Fig. 7. Unlike DCT-Corr [48] [Fig. 7(b)], our fused image [Fig. 7(e)] contains no visible artifacts around the edges where the focus transition happens and presents a more natural output. Compared to the reference CNN-based methods [58] and [63], the proposed model generates a sharper all-in-focus image. As can be seen, the numbers on the shirts of players 9 and 23 (standing close to the net) look sharper in Fig. 7(e) than in Fig. 7(c) and Fig. 7(d) (best viewed on the digital version).

2) Multi-exposure image fusion: In multi-exposure image fusion, we aim at reconstructing a photo-realistic image by fusing under-exposed and over-exposed images. We utilize the proposed CNN with a recently published multi-exposure dataset [80] which contains seven exposure levels for each scene. We obtain a training dataset with $30,000$ pairs of image patches, including the first level as the under-exposed image and the sixth level as the over-exposed image. Since the first level corresponds to a very dark image and the sixth level to a very bright image, providing the desired photo-realistic image is a challenging task. As a reference method we use CU-Net [25].

Figures 5 and 6 present visual results for two example images, namely, “tree” and “church”. For the “tree” image,

TABLE III
MULTIMODAL DENOISING OF FLASH/NON-FLASH IMAGES. PERFORMANCE COMPARISON [IN TERMS OF PSNR (DB)] FOR SELECTED TEST IMAGES AT THREE DIFFERENT NOISE LEVELS.

Flash/non-flash		Minion	Towel	Elmo	Pendant	Book	Tampax	Typewriter	Pot	Plant	Flower	Aloe	Cactus	Average
$\sigma = 25$	CBM3D [77]	33.63	37.25	36.03	39.39	36.17	35.94	34.37	33.91	33.82	35.62	31.86	31.38	34.95
	DnCNN [78]	34.13	37.58	36.65	39.78	35.61	36.55	34.87	34.35	34.42	36.26	32.70	31.62	35.38
	DJF [39]	31.59	36.86	34.69	36.83	34.16	34.45	32.81	33.26	31.90	34.69	30.78	31.13	33.76
	MuGIF [31]	30.49	35.42	33.75	35.78	32.62	33.46	31.51	31.82	30.95	33.24	29.49	30.88	32.45
	CU-Net [25]	34.24	37.99	36.82	39.95	36.86	36.97	35.07	35.52	34.42	36.40	32.83	33.26	35.86
	proposed	**34.52**	**38.43**	**37.05**	**40.26**	**37.10**	**37.22**	**35.54**	**35.95**	**34.74**	**36.91**	**33.07**	**33.62**	**36.21**
$\sigma = 50$	CBM3D [77]	29.94	34.36	32.50	36.40	33.15	32.37	31.30	30.57	30.17	32.33	27.98	27.83	31.58
	DnCNN [78]	30.32	34.65	33.24	36.58	32.37	32.95	31.87	30.87	30.83	33.04	28.49	28.04	31.94
	DJF [39]	28.55	33.27	31.79	34.65	31.74	31.58	29.90	30.02	28.84	31.30	28.05	27.66	30.61
	MuGIF [31]	26.93	32.02	30.94	31.78	29.72	30.54	28.57	28.97	27.45	29.93	28.97	27.81	29.22
	CU-Net [25]	31.08	35.91	34.20	37.23	34.11	34.22	32.40	32.88	31.16	33.60	29.17	30.69	33.05
	proposed	**31.32**	**36.07**	**34.36**	**37.42**	**34.41**	**34.41**	**32.70**	**33.06**	**31.48**	**33.81**	**29.34**	**30.88**	**33.28**
$\sigma = 75$	CBM3D [77]	27.85	32.34	30.75	34.41	31.40	30.31	29.56	28.79	28.20	30.40	25.83	26.10	29.66
	DnCNN [78]	28.33	32.75	31.16	34.82	31.83	30.74	29.97	29.21	28.46	30.83	26.25	26.55	30.08
	DJF [39]	26.65	31.77	30.63	32.78	29.54	29.86	28.85	28.33	27.54	29.25	25.56	26.26	28.92
	MuGIF [31]	24.89	30.30	29.82	29.70	28.40	28.54	26.46	27.44	25.82	27.84	24.49	26.27	27.50
	CU-Net [25]	29.12	34.25	32.55	35.12	32.93	32.19	30.55	31.66	29.17	31.73	27.03	29.30	31.30
	proposed	**29.39**	**34.48**	**32.75**	**35.44**	**33.19**	**32.44**	**30.89**	**31.85**	**29.47**	**32.03**	**27.24**	**29.56**	**31.56**



(a) SRFBN [73] (b) DJF [39] (c) CoISTA [21] (d) LMCSC [24] (e) CU-Net [25] (f) proposed

Fig. 3. $\times 4$ super-resolution of the multi-spectral image "chart toy" and the corresponding error maps. The proposed model is compared against a single-modal (a) and four multimodal SR methods (b)-(e). The high quality reconstruction achieved by the proposed model in (f) results in a PSNR equal to 41.13. The PSNR values for the second (d) and the third (e) best methods are 40.31 and 39.47, respectively. The superior performance of the proposed method is more clear in the upper part of the image which contains structural details like numbers and shapes, implying a more efficient use of the side information. All the multimodal methods (b)-(f) outperform the single-modal technique (a).

TABLE IV
DIFFERENT REALIZATIONS OF THE PROPOSED CNN WITH VARYING FILTER SIZES OF THE CONVOLUTIONAL DICTIONARY. RESULTS ARE PRESENTED IN TERMS OF AVERAGE PSNR FOR THE SUPER-RESOLUTION OF MULTISPECTRAL IMAGES AT SCALE 8.

filter size	3×3	5×5	7×7	9×9
PSNR	36.07	36.23	36.47	36.38

TABLE V
COMPARISON OF THE INFERENCE TIME (SEC) OF THE PROPOSED MODEL AGAINST SEVERAL REFERENCE METHODS W.R.T. TWO DIFFERENT INPUTS.

input size	SDF [30]	DGF [40]	LMCSC [24]	CU-Net [25]	proposed
256×256	3.12	1.78	1.09	0.96	1.38
512×512	3.63	2.07	1.29	1.14	1.61

the proposed fusion model [Fig. 5(e)] results in a more natural output with less halo-like artifacts around the tree compared to CU-Net [25] [Fig. 5(d)]. For the "church" image, we can see that the result of CU-Net [Fig. 6(d)] contains some shadow-like artifacts around the window, and the art designs on the window glasses are blurry, while the proposed model [Fig. 6(e)] results in a sharper and more natural image.

D. Ablation study

In this section, we study the effect of different network parameters—including the number of C-CSC stages, the number of iterations implemented at each C-CSC stage and the filter size of the convolutional dictionary—on the performance of the proposed network. The comparison is performed using the Columbia multi-spectral database[4] for the task of multimodal

[4]http://www.cs.columbia.edu/CAVE/databases/multispectral

TABLE VI

DIFFERENT REALIZATIONS OF THE PROPOSED CNN WITH A VARYING NUMBER OF C-CSC STAGES AND IMPLEMENTED ITERATIONS AT EACH STAGE. RESULTS ARE PRESENTED IN TERMS OF AVERAGE PSNR FOR THE SUPER-RESOLUTION OF MULTISPECTRAL IMAGES AT SCALE 8. THE TABLE ALSO INCLUDES THE CORRESPONDING NETWORK COMPLEXITY [IN TERMS OF NUMBER OF FLOPS ($\times 10^{11}$)] FOR EACH CONFIGURATION.

MS/RGB $\times 8$ SR	#stages = 1		#stages = 2		#stages = 3		#stages = 4	
	PSNR (dB)	FLOPs ($\times 10^{11}$)	PSNR (dB)	FLOPs ($\times 10^{11}$)	PSNR (dB)	FLOPs ($\times 10^{11}$)	PSNR (dB)	FLOPs ($\times 10^{11}$)
#iterations = 1	35.73	0.2164	36.14	0.4287	**36.47**	**0.6411**	36.51	0.8534
#iterations = 2	35.84	0.3488	36.19	0.6935	36.49	1.0382	36.55	1.3829
#iterations = 3	35.89	0.4812	36.20	0.9582	36.52	1.4353	36.59	1.9124

TABLE VII

PERFORMANCE COMPARISON BETWEEN THE PROPOSED BIDIRECTIONAL AND A SIMILAR SINGLE-DIRECTIONAL NETWORK ARCHITECTURE. RESULTS ARE PRESENTED IN TERMS OF AVERAGE PSNR AND SSIM FOR THE SUPER-RESOLUTION OF MULTISPECTRAL IMAGES AT SCALES 4 AND 8.

MS/RGB	Chart toy		Egyptian		Feathers		Glass tiles		Jelly beans		Oil Paintings		Paints		Average	
$\times 4$	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
modified	40.28	0.9968	49.02	0.9985	41.53	0.9961	34.57	0.9941	39.87	0.9969	39.19	0.9917	39.11	0.9967	40.51	0.9958
proposed	**41.13**	**0.9977**	**49.92**	**0.9991**	**42.27**	**0.9974**	**34.84**	**0.9948**	**40.75**	**0.9980**	**39.84**	**0.9926**	**39.69**	**0.9978**	**41.20**	**0.9968**
$\times 8$	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
modified	34.52	0.9812	43.88	0.9962	36.94	0.9882	30.13	0.9734	34.65	0.9887	36.31	0.9763	34.76	0.9910	35.88	0.9851
proposed	**35.47**	**0.9837**	**44.40**	**0.9969**	**37.83**	**0.9901**	**30.89**	**0.9797**	**34.79**	**0.9890**	**36.98**	**0.9793**	**34.96**	**0.9927**	**36.47**	**0.9873**

image SR by a factor of 8. The experiments are performed on the same multispectral test set employed in Section VI-B and results are presented in terms of average PSNR.

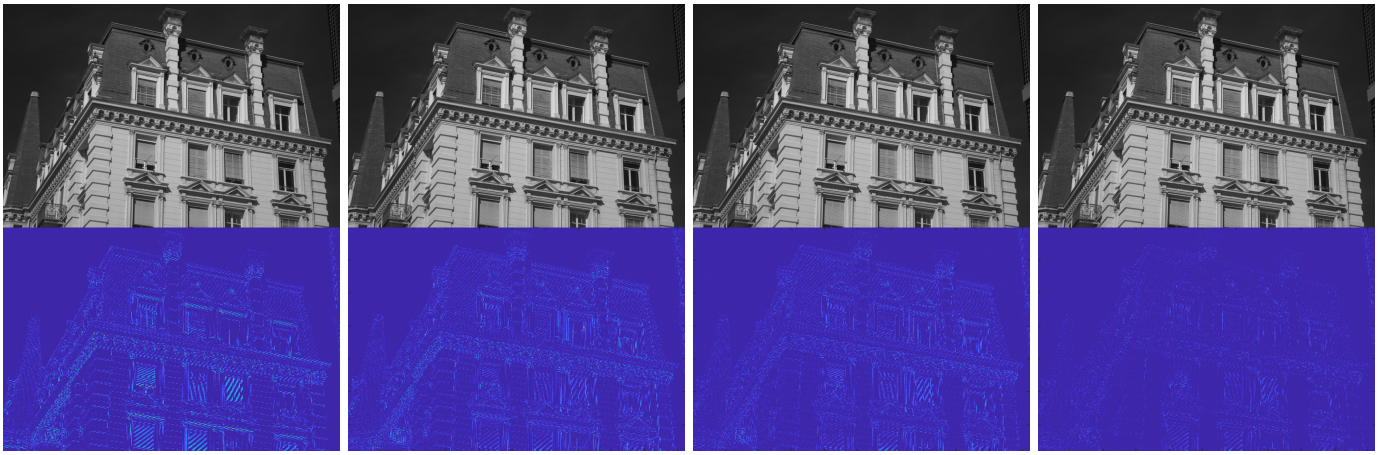
First, we evaluate the proposed multimodal restoration network with three C-CSC stages each implementing one iteration, and four different filter sizes, i.e., 3×3, 5×5, 7×7 and 9×9. The results depicted in Table IV suggest an increase in performance for the size of 7×7. We fix the size of the dictionary filters to 7×7 and provide the results on the same task using a network with different number of stages $(1, 2, 3, 4)$ and different number of iterations $(1, 2, 3)$. We summarize the results of this ablation study in Table VI. Table VI also includes results for the complexity of the corresponding network expressed in terms of floating point operations (FLOPs); the number of FLOPs is computed for a forward pass of the trained network for input images of size 512×512. The bold numbers in the table correspond to the network design used for the experiments presented in Section VI-B. We observe that increasing both the number of C-CSC stages and the number of implemented iterations at each stage improves the performance; however, the PSNR gain from adding C-CSC stages is more significant. By taking into account both the performance gain and the complexity, an interesting observation is that a higher complexity does not always lead to a better performance. For instance, a design with three C-CSC stages and a single sparse coding iteration outperforms a network design with two stages and two or three iterations while having a lower complexity. If we use only one stage with three sparse coding iterations, the design is similar to our previous LMCSC-Net presented in [24].

Concerning the complexity of the proposed multimodal CNN compared to other multimodal deep learning designs, we report the following: The proposed CNN configuration used in the experiments presented in Sections VI-B, VI-C contains 112K learnable parameters, a number similar to or lower than the number of parameters in alternative best competing methods, i.e., LMCSC-Net [24] with 98K parameters, CU-Net [25] with 151K parameters, and CoISTA [21] with 818K parameters. A comparison with respect to the inference time is presented in Table V. All models are tested on a machine with an NVIDIA GeForce GTX 1070 GPU. Note that our design outperforms these models on different tasks.

Finally, the last set of experiments includes a comparison of the proposed bidirectional network design, which alternates the guidance role of the input modalities, with a similar single-directional network that keeps the guidance representation fixed and only updates the representations of the target modality (similar to LMCSC-Net in [24]). We build such a network by replacing the LeSITA operator ξ_{μ_2} used in the beta updating blocks (see (29)) with the simple soft thresholding operator $\phi_\gamma(\mathbf{u}) = \text{sign}(\mathbf{u}) \max(0, |\mathbf{u}| - \gamma)$. The hyper-parameters of both networks are the same. Table VII presents the results of these two networks for multispectral image SR upscaling $\times 4$, $\times 8$. The results clearly show the superior performance of the proposed design, indicating the significant role of the LeSITA operator in the fusion process.

VII. CONCLUSION

In this paper, we presented a deep unfolding CNN design for coupled convolutional sparse coding that relies on the method of multipliers. The proposed CNN unfolds several stages of the numerical algorithm performing coupled encoding of the multimodal input data. For image reconstruction tasks, a final reconstruction/fusion layer is added to generate the output image. We have built two multimodal models tailored to multimodal image restoration and image fusion. We have provided an ablation study where the parameterization of the network is investigated experimentally. The superior performance of the proposed design against existing single-modal and multimodal designs was demonstrated by experimental results on several multimodal datasets employed for multimodal image restoration and image fusion. The provided numerical and



(a) FSRCNN [14] (b) CoISTA [21] (c) LMCSC [24] (d) proposed

Fig. 4. $\times 2$ super-resolution of the NIR image "o-0018" and the corresponding error maps. The proposed model is compared against a single modal SR method in (a) and two multimodal deep unfolding designs in (b), (c), using RGB images as side information. Specifically, the results in (c) are obtained by our previous work [24]. The error maps highlight an improved reconstruction of the structural details by the proposed model and show the ability of the proposed architecture to integrate useful structural information from the RGB modality.

(a) under exposed (b) over exposed (c) ground truth (d) CU-Net [25] (e) proposed

Fig. 5. A multi-exposure image fusion example for image "tree". Reconstruction with the proposed fusion model in (e) is compared against the multimodal CU-Net [25] in (d). Both images are of high quality, however, the proposed design achieves a more natural output with less halo-like artifacts around the tree.

(a) under exposed (b) over exposed (c) ground truth (d) CU-Net [25] (e) proposed

Fig. 6. A multi-exposure image fusion example for image "church". Reconstruction with the proposed fusion model in (e) is compared against the multimodal CU-Net [25] in (d). The image obtained with CU-Net contains shadow-like artifacts around the window and blurriness in the art designs on the window glasses, whereas the proposed model computes a sharper and more natural image.

(a) Source A, B (b) DCT-Corr [48] (c) [58] (d) [63] (e) proposed

Fig. 7. A multi-focus image fusion example. The proposed model in (e) is compared against a transform-domain method in (b), and two CNN designs in (c), (d). The proposed design generates a sharper all-in-focus image with no visible artifacts around edges, resulting in a more natural output.

visual results corroborate that the proposed network structure incorporates domain knowledge from sparse representations efficiently, allowing fusion both at the encoding and decoding steps, and the trained models offer accurate and fast inference. Our model can be employed in other multimodal restoration tasks that can be formulated as a linear inverse problem such as compressive image reconstruction or image inpainting. We will address these applications in our future work.

REFERENCES

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

[2] 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.

[3] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Information Fusion*, vol. 40, pp. 57–75, 2018.

[4] 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.

[5] I. Tosic and S. Drewes, "Learning joint intensity-depth sparse representations," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2122–2132, 2014.

[6] N. Deligiannis, J. F. C. Mota, B. Cornelis, M. R. D. 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.

[7] N. Deligiannis, J. F. Mota, B. Cornelis, M. R. Rodrigues, and I. Daubechies, "X-ray image separation via coupled dictionary learning," in *IEEE International Conference on Image Processing (ICIP)*, 2016.

[8] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis, P. L. Dragotti, and M. R. D. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Transactions on Computational Imaging*, 2019.

[9] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 884–892, 2009.

[10] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, pp. 72–84, 2015.

[11] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information Fusion*, vol. 24, pp. 147–164, 2015.

[12] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 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 *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, 2016, pp. 391–407.

[15] C. Xie, W. Zeng, and X. Lu, "Fast single-image super-resolution via deep network with component learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3473–3486, 2019.

[16] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[17] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *IEEE International Conference on Machine Learning (ICML)*, 2010.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] ——, "Multimodal deep unfolding for guided image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 8443–8456, 2020.

[25] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[26] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Designing CNNs for multimodal image super-resolution via the method of multipliers," in *European Signal Processing Conference (EUSIPCO)*, 2020.

[27] S.-W. Jung and O. Choi, "Learning-based filter selection scheme for depth image super resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1641–1650, 2014.

[28] 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.

[29] 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.

[30] 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.

[31] X. Guo, Y. Li, J. Ma, and H. Ling, "Mutually guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 694–707, 2020.

[32] X. Zhang, T. Sim, and X. Miao, "Enhancing photographs with near infra-red images," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[33] M. Kiechle, S. Hawe, and M. Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1545–1552.

[34] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," in *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 2012, pp. 152–157.

[35] X. Deng, P. Song, M. R. Rodrigues, and P. L. Dragotti, "RADAR: Robust algorithm for depth image super resolution based on FRI theory and multimodal dictionary learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2447–2462, 2019.

[36] P. Song and M. R. D. Rodrigues, "Multimodal image denoising based on coupled dictionary learning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 515–519.

[37] 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.

[38] H. Kwon and Y. Tai, "RGB-Guided Hyperspectral Image Upsampling," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 307–315.

[39] Y. Li, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep joint image filtering," in *European Conference on Computer Vision (ECCV)*, 2016.

[40] 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.

[41] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Joint image super-resolution via recurrent convolutional neural networks with coupled sparse priors," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 868–872.

[42] H. Sreter and R. Giryes, "Learned convolutional sparse coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[43] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer graphics forum*, vol. 28, no. 1. Wiley Online Library, 2009, pp. 161–171.

[44] A. A. Goshtasby, "Fusion of multi-exposure images," *Image and Vision Computing*, vol. 23, no. 6, pp. 611–618, 2005.

[45] E. S. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–12.

[46] S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 626–632, 2012.

[47] H. Li, T. N. Chan, X. Qi, and W. Xie, "Detail-Preserving Multi-Exposure Fusion with Edge-Preserving Structural Patch Decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[48] M. Amin-Naji and A. Aghagolzadeh, "Multi-focus image fusion in DCT domain using variance and energy of laplacian and correlation coefficient for visual sensor networks," *Journal of AI and Data Mining*, vol. 6, no. 2, pp. 233–250, 2018.

[49] M. Abdipour and M. Nooshyar, "Multi-focus image fusion using sharpness criteria for visual sensor networks in wavelet domain," *Computers & Electrical Engineering*, vol. 51, pp. 74–88, 2016.

[50] B. Yu, B. Jia, L. Ding, Z. Cai, Q. Wu, R. Law, J. Huang, L. Song, and S. Fu, "Hybrid dual-tree complex wavelet transform and support vector machine for digital multi-focus image fusion," *Neurocomputing*, vol. 182, pp. 1–9, 2016.

[51] S. Liu, J. Chen, and S. Rahardja, "A new multi-focus image fusion algorithm and its efficient implementation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 5, pp. 1374–1384, 2019.

[52] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1037–1041, 2015.

[53] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[54] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[55] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, p. 594, 2016.

[56] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sensing and Imaging*, vol. 17, no. 1, pp. 1–16, 2016.

[57] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.

[58] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[59] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4724–4732.

[60] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1102–1115, 2018.

[61] H. Li and X. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[62] R. Nie, J. Cao, D. Zhou, and W. Qian, "Multi-source information exchange encoding with PCNN for medical image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 986–1000, 2020.

[63] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, and D. Zhang, "DRPL: Deep regression pair learning for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4816–4831, 2020.

[64] M. Elad, *Sparse and redundant representations: From theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

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

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

[67] L. Weizman, Y. C. Eldar, and D. B. Bashat, "Compressed sensing for longitudinal MRI: An adaptive-weighted approach," *Medical Physics*, vol. 42, no. 9, pp. 5195–5208, 2015.

[68] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.

[69] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, 2012.

[70] V. Stanković, L. Stanković, and S. Cheng, "Compressive image sampling with side information," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 3037–3040.

[71] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," 2005.

[72] M. Trocan, T. Maugey, J. E. Fowler, and B. Pesquet-Popescu, "Disparity-compensated compressed-sensing reconstruction for multiview images," in *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 2010, pp. 1225–1229.

[73] 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.

[74] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, "Closed-loop matters: Dual regression networks for single image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[75] 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.

[76] Y. Aksoy, C. Kim, P. Kellnhofer, S. Paris, M. Elgharib, M. Pollefeys, and W. Matusik, "A dataset of flash and ambient illumination pairs from the crowd," in *ECCV*, 2018.

[77] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image restoration by sparse 3D transform-domain collaborative filtering," in *Image Processing: Algorithms and Systems VI*, J. T. Astola, K. O. Egiazarian, and E. R. Dougherty, Eds., vol. 6812, International Society for Optics and Photonics. SPIE, 2008, pp. 62 – 73.

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

[79] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, pp. 72 – 84, 2015.

[80] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.