Designing CNNs for Multimodal Image Super-Resolution via the Method of Multipliers

Iman Marivani, Evaggelia Tsiligianni, Bruno Cornelis, Nikos Deligiannis
Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
imec, Kapeldreef 75, B-3001 Leuven, Belgium

***Abstract*—Multimodal alias, guided, image super-resolution (SR) refers to the reconstruction of a high-resolution (HR) version of a low-resolution (LR) image with the aid of an HR image from another image modality. Common approaches for the SR problem include analytical methods which are computationally expensive. Deep learning methods are capable of learning a nonlinear mapping between LR and HR images from data, delivering high reconstruction accuracy at a low-computational cost during inference; however, these methods do not incorporate any prior knowledge about the problem, with the neural network model behaving like a black box. In this paper, we formulate multimodal image SR as a coupled convolutional sparse coding problem. To solve the corresponding minimization problem, we adopt the Method of Multipliers (MM). We then design a convolutional neural network (CNN) that unfolds the obtained MM algorithm. The proposed CNN accepts as input the LR image from the main modality and the HR image from the guidance modality to reconstruct the desired HR image. Unlike existing deep learning methods, our CNN provides an efficient and structured way to fuse information at different stages of the network and achieves high reconstruction accuracy. We evaluate the performance of the proposed model for the super-resolution of multi-spectral images guided by their high resolution RGB counterparts.**

***Index Terms*—Method of multipliers, deep unfolding, multimodal image super-resolution, multimodal CNN.**

I. Introduction

Image super-resolution (SR) refers to the restoration of a high-resolution (HR) image given a low-resolution (LR) observation. In many applications, an image from another modality can be used as reference [1]–[4]. Multimodal image SR is the reconstruction of an HR image from its LR version, guided by another image modality [1], [5]. Several studies [6]–[8] have addressed image SR via sparse coding. These methods are based on the assumption that the LR and HR images share the same sparse representation w.r.t. different dictionaries, and rely on analytical approaches to solve complex optimization problems both at training and inference. A sparsity-based solution for multimodal image SR has been presented in [1]. Deep learning methods have significantly enhanced the performance of image SR, by learning a nonlinear mapping between the LR/HR images from training data [9]–[13]. While these models deliver high performance as well as fast inference, their black-box design does not allow integration of domain knowledge about the problem and lack a theoretical foundation.

Deep unfolding [14], [15] introduced the idea of integrating prior knowledge, e.g., signal sparsity, into the neural network architecture. Existing designs consist of layers performing operations similar to iterative algorithms for sparse coding [14], [15]. Single-modal deep unfolding networks have been employed for image SR [16], image denoising [17] and compressive sensing MRI [15]; the model in [15] is a single-modal deep unfolding design that relies on the Alternating Direction Method of Multipliers (ADMM). Recently, multimodal deep unfolding for image SR has been presented in [3], [4], [18]. The authors of [18] proposed a network that combines two branches, which perform sparse coding of the target and the guidance images. The two modalities are fused at a final layer by linearly combining the respective sparse representations. Different from [18], our previous work [3], [4], [19] is inspired by proximal methods for sparse approximation with side information, and proposes a multimodal architecture with several layers, each one performing fusion of the two modalities.

In this paper, we address multimodal image SR as a coupled convolutional sparse coding problem, and propose a deep unfolding solution based on the method of multipliers (MM). Specifically, we formulate coupled convolutional sparse coding as a constrained optimization problem, we exploit alternating minimization method to obtain two sub-problems regarding the two modalities, and utilize MM to obtain an iterative algorithm which computes sparse representations of the image modalities. The algorithm is then translated into a multimodal convolutional neural network. The network accepts as input the LR image from the target modality and the HR image from the guidance modality, and consists of a few stages each corresponding to a single iteration of the MM method. At every stage, the model jointly learns sparse representations of both modalities. Finally the desired HR image of the target modality is reconstructed using a convolutional dictionary layer. Similar to our previous work [3], [4], the proposed model performs fusion of information at every stage; however, the design differs from [3], [4] as the intermediate representations of the two modalities are learned at every stage, while in [3], [4] the representation of the guidance modality is provided by a second network branch performing sparse coding. The proposed model is employed to super-resolve multi-spectral images with the aid of high resolution RGB images. Experimental results show a significant performance improvement compared to the state of the art.

The paper is organized as follows. Section II includes the necessary background. In Section III, we present the proposed approach. Experimental results are presented in Section IV,

This research received funding from the FWO (Project G0A2617N) and by Innoviris (Project ROADMAP), Belgium.

and conclusions are drawn in Section V.

II. 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) [20]. 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 [21]. CSC is formulated as follows:

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

where $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ represents the input image, $\mathbf{D}_i \in \mathbb{R}^{p_1 \times p_2}$, $i = 1, ..., k$, are the atoms of a convolutional dictionary $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{p_1 \times p_2 \times k}$, and $\mathbf{A}_i \in \mathbb{R}^{n_1 \times n_2}$, $i = 1, ..., k$, 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. Method of Multipliers

MM [22] is an efficient algorithm for the solution of constrained optimization problems of the form

$$\begin{aligned} &\min_{\mathbf{p}} \ f(\mathbf{p}) \\ &\text{s.t. } \mathbf{A}\mathbf{p} = \mathbf{c}. \end{aligned} \tag{3}$$

The algorithm involves the solution of the $\mathbf{p}$ sub-problem by minimizing the augmented Lagrangian function. Let us define the augmented Lagrangian function as

$$L(\mathbf{p}, \boldsymbol{\rho}) = f(\mathbf{p}) + \langle \boldsymbol{\rho}, \mathbf{A}\mathbf{p} - \mathbf{c} \rangle + \frac{\eta}{2}\|\mathbf{A}\mathbf{p} - \mathbf{c}\|_2^2, \tag{4}$$

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:

$$\begin{cases} \mathbf{p}^{k+1} = \arg\min_{\mathbf{p}} L(\mathbf{p}, \boldsymbol{\rho}^k), \\ \boldsymbol{\rho}^{k+1} = \boldsymbol{\rho}^k + \eta(\mathbf{A}\mathbf{p}^{k+1} - \mathbf{c}). \end{cases} \tag{5}$$

Depending on the problem at hand, various methods can be used for the solution of the minimization problem in (5). Next, we discuss proximal methods.

C. Proximal methods

Proximal methods [23] have been proposed for the solution of optimization problems of the form

$$\min_{\mathbf{p}} \ h(\mathbf{p}) + \lambda g(\mathbf{p}), \tag{6}$$

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

$$\mathbf{p}^{k+1} = \text{Prox}_{\mu}(\mathbf{p}^k - \frac{1}{L}\nabla h(\mathbf{p}^k)), \tag{7}$$

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

$$\text{Prox}_{\mu}(\mathbf{u}) = \arg\min_{\mathbf{v}} \{\frac{1}{2}\|\mathbf{v} - \mathbf{u}\|_2^2 + \mu g(\mathbf{v})\}. \tag{8}$$

III. Proposed method

In this paper, we address the problem of finding a target HR image from its LR version with the aid of an HR image coming from a guidance image modality. Our approach relies on the assumption that the LR and HR images from the target modality share the same sparse representations under different dictionaries; therefore, finding the HR image is equivalent to computing the sparse codes of the LR image [6]. However, computing a sparse representation of the LR image might not be an easy task due to blurring artifacts and noise in the data. In order to accurately estimate the desired representation, we employ an HR guidance image under the assumption that the sparse representations of the target image and the guidance HR image are similar by means of the ℓ_1-norm. This way multimodal image super-resolution can be addressed as a coupled sparse coding problem. In what follows, we present the proposed approach in detail.

A. Coupled Sparse Coding via the Method of Multipliers

We denote by $\mathbf{y} \in \mathbb{R}^n$ a vectorized LR patch from the target modality, and $\mathbf{z} \in \mathbb{R}^n$ the corresponding HR patch from the guidance modality. We assume that $\mathbf{y}$ and $\mathbf{z}$ have sparse codes $\boldsymbol{\alpha} \in \mathbb{R}^m$, $\boldsymbol{\beta} \in \mathbb{R}^m$ under dictionaries $\mathbf{D}_y$, $\mathbf{D}_z$, respectively. It is also assumed that the underlying correlation between the modalities can be captured using an ℓ_1 term [24], [25]. Then, the coupled sparse coding problem can be formulated in the form of a constrained minimization problem as

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \ \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 \\ &\text{s.t.} \quad \mathbf{D}_y\boldsymbol{\alpha} = \mathbf{y}, \quad \mathbf{D}_z\boldsymbol{\beta} = \mathbf{z}, \end{aligned} \tag{9}$$

with κ a regularization parameter. While the first two ℓ_1 terms of (9) concern the sparse coding of the image modalities, the last term captures the similarities between them. The objective in (9) is convex w.r.t. $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ while the other is kept fixed. Therefore, we propose solving (9) as two sub-problems regarding the sparse representations $\boldsymbol{\alpha}$ when $\boldsymbol{\beta}$ is kept fixed and vice versa, alternatively. We solve the obtained sub-problems using the MM method. The augmented Lagrangian function for the sub-problem concerning $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\rho}_1) = \|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}^k\|_1 + \langle \boldsymbol{\rho}_1, \mathbf{D}_y\boldsymbol{\alpha} - \mathbf{y} \rangle \\ + \frac{\eta_1}{2}\|\mathbf{D}_y\boldsymbol{\alpha} - \mathbf{y}\|_2^2, \end{aligned} \tag{10}$$

and similarly for $\boldsymbol{\beta}$ we have

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\rho}_2) = \|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\beta}\|_1 + \langle \boldsymbol{\rho}_2, \mathbf{D}_z\boldsymbol{\beta} - \mathbf{z} \rangle \\ + \frac{\eta_2}{2}\|\mathbf{D}_z\boldsymbol{\beta} - \mathbf{z}\|_2^2, \end{aligned} \tag{11}$$

with $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$ the corresponding Lagrange multipliers and η_1, η_2 regularization parameters. The sparse code $\boldsymbol{\alpha}$, and the Lagrange parameter $\boldsymbol{\rho}_1$ are updated as follows:

$$\boldsymbol{\alpha}^{k+1} = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}^k\|_1 + \langle \boldsymbol{\rho}_1^k, \mathbf{D}_y\boldsymbol{\alpha} - \mathbf{y}\rangle + \frac{\eta_1}{2}\|\mathbf{D}_y\boldsymbol{\alpha} - \mathbf{y}\|_2^2\}, \tag{12}$$

$$\boldsymbol{\rho}_1^{k+1} = \boldsymbol{\rho}_1^k + \eta_1(\mathbf{D}_y\boldsymbol{\alpha}^{k+1} - \mathbf{y}), \tag{13}$$

Additionally, we have the update of the side information sparse code $\boldsymbol{\beta}$, and the corresponding Lagrange parameter $\boldsymbol{\rho}_2$ as:

$$\boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}} \{\|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\beta}\|_1 + \langle \boldsymbol{\rho}_2^k, \mathbf{D}_z\boldsymbol{\beta} - \mathbf{z}\rangle + \frac{\eta_2}{2}\|\mathbf{D}_z\boldsymbol{\beta} - \mathbf{z}\|_2^2\}, \tag{14}$$

$$\boldsymbol{\rho}_2^{k+1} = \boldsymbol{\rho}_2^k + \gamma_2(\mathbf{D}_z\boldsymbol{\beta}^{k+1} - \mathbf{z}). \tag{15}$$

Problems (12) and (14) can be solved with proximal methods. By setting $h(\boldsymbol{\alpha}) = \langle \boldsymbol{\rho}_1^k, \mathbf{D}_y\boldsymbol{\alpha} - \mathbf{y}\rangle + \frac{\eta_1}{2}\|\mathbf{D}_y\boldsymbol{\alpha} - \mathbf{y}\|_2^2$, $g(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 + \kappa\|\boldsymbol{\alpha} - \boldsymbol{\beta}^k\|_1$, $\lambda = 1$, the proximal operator is obtained from

$$\xi_\mu(\mathbf{u}) = \arg\min_{\mathbf{v}} \Big\{\frac{1}{2}\|\mathbf{v}-\mathbf{u}\|_2^2 + \mu(\|\mathbf{v}\|_1 + \kappa\|\mathbf{v} - \boldsymbol{\beta}^k\|_1)\Big\}. \tag{16}$$

In [19] an ℓ_1-ℓ_1 minimization problem similar to (12) was addressed using proximal methods. Both problems involve a non-smooth term of the same form, that is, a sum of ℓ_1-norms. Since the proximal operator depends only on this term, it can be easily shown that ξ_μ has the same form with the proximal operator presented in [19] with parameter $\mu_1 = \mu(1+\kappa)/2$. Therefore, the proximal algorithm for the update of $\boldsymbol{\alpha}$ takes the form

$$\boldsymbol{\alpha}^{k+1} = \xi_{\mu_1}\big(\boldsymbol{\alpha}^k - \frac{1}{L}(\mathbf{D}_y^T\boldsymbol{\rho}_1^k + \eta_1\mathbf{D}_y^T\mathbf{D}_y\boldsymbol{\alpha}^k - \eta_1\mathbf{D}_y^T\mathbf{y})\big). \tag{17}$$

We address the problem (14) in a similar way, by formulating a proximal solution for the update of $\boldsymbol{\beta}$ given by

$$\boldsymbol{\beta}^{k+1} = \xi_{\mu_2}\big(\boldsymbol{\beta}^k - \frac{1}{L}(\mathbf{D}_z^T\boldsymbol{\rho}_2^k + \eta_2\mathbf{D}_z^T\mathbf{D}_z\boldsymbol{\beta}^k - \eta_2\mathbf{D}_z^T\mathbf{z})\big), \tag{18}$$

with $\mu_2 = \mu(1+\kappa)/2$. Note that we use different notations for the parameters of the proximal operators in (17), (18) because these equations will guide the design of a learnable model which will learn μ_1, μ_2.

B. Multimodal CNN design

In order to apply the proposed solution to entire images instead of image patches, we can rewrite the objective (9) in the form of convolutional sparse coding, that is,

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}}} \sum_{i=1}^k \|\mathbf{A}_i\|_1 + \sum_{i=1}^k \|\mathbf{B}_i\|_1 + \kappa\sum_{i=1}^k \|\mathbf{A}_i - \mathbf{B}_i\|_1 \\ &\text{s.t.} \quad \sum_{i=1}^k \mathbf{D}_i^Y * \mathbf{A}_i = \mathbf{Y}, \quad \sum_{i=1}^k \mathbf{D}_i^Z * \mathbf{B}_i = \mathbf{Z}, \end{aligned} \tag{19}$$

where $\mathbf{Y} \in \mathbb{R}^{n_1\times n_2}$ represents the input LR image and $\mathbf{Z} \in \mathbb{R}^{n_1\times n_2}$ is the guidance HR image; $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times n_2\times k}$, $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_1\times n_2\times k}$, are the corresponding sparse feature maps

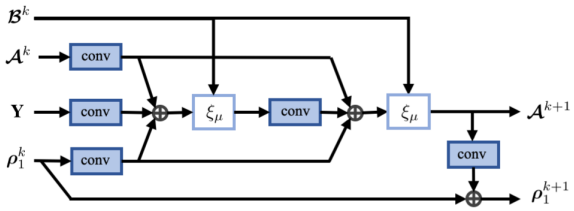


Fig. 1: Unfolding the update of $\boldsymbol{\mathcal{A}}$ performed by a single iteration of MM for coupled CSC. The depicted sub-network also shows the update of the parameter $\boldsymbol{\rho_1}$. The sub-network for the update of $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\rho_2}$ has a similar structure and accepts as input the guidance image $\mathbf{Z}$ instead of the main image $\mathbf{Y}$.

with respect to convolutional dictionaries $\boldsymbol{\mathcal{D}}^Y$, $\boldsymbol{\mathcal{D}}^Z$, with atoms $\mathbf{D}_i^Y \in \mathbb{R}^{p_1\times p_2}$, $\mathbf{D}_i^Z \in \mathbb{R}^{p_1\times p_2}$, $i = 1, ..., k$, respectively. Considering the linear properties of convolution, the convolutional terms in (19) can be replaced by matrix-vector multiplications (as in conventional sparse coding). Specifically, the convolutional term $\sum_{i=1}^k \mathbf{D}_i^Y * \mathbf{A}_i$ can be replaced by a linear term $\boldsymbol{\Phi}_y\boldsymbol{\alpha}$, with $\boldsymbol{\Phi}_y \in \mathbb{R}^{(n_1-p_1+1)(n_2-p_2+1)\times kn_1n_2}$ a matrix with Toeplitz structure, and $\boldsymbol{\alpha} \in \mathbb{R}^{kn_1n_2}$ containing the vectorized sparse feature maps $\boldsymbol{\mathcal{A}}$ of the target modality. Similarly, we set $\boldsymbol{\Phi}_z\boldsymbol{\beta} := \sum_{i=1}^k \mathbf{D}_i^Z * \mathbf{B}_i$. Therefore, problem (19) takes a form similar to (9) and can be addressed with the proposed MM solution. At each iteration the unknown convolutional sparse codes are updated as follows:

$$\boldsymbol{\alpha}^{k+1} = \xi_{\mu_1}\big(\boldsymbol{\alpha}^k - \frac{1}{L}(\boldsymbol{\Phi}_y^T\boldsymbol{\rho}_1^k + \eta_1\boldsymbol{\Phi}_y^T\boldsymbol{\Phi}_y\boldsymbol{\alpha}^k - \eta_1\boldsymbol{\Phi}_y^T\mathbf{y})\big), \tag{20}$$

$$\boldsymbol{\beta}^{k+1} = \xi_{\mu_2}\big(\boldsymbol{\beta}^k - \frac{1}{L}(\boldsymbol{\Phi}_z^T\boldsymbol{\rho}_2^k + \eta_2\boldsymbol{\Phi}_z^T\boldsymbol{\Phi}_z\boldsymbol{\beta}^k - \eta_2\boldsymbol{\Phi}_z^T\mathbf{z})\big). \tag{21}$$

Taking into account the structure of matrices $\boldsymbol{\Phi}_y$ and $\boldsymbol{\Phi}_z$, we can rewrite (20), (21) in the form of convolutions, that is,

$$\boldsymbol{\mathcal{A}}^{k+1} = \xi_{\mu_1}\big(\boldsymbol{\mathcal{A}}^k - \boldsymbol{\mathcal{Q}} * \boldsymbol{\rho}_1^k + \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{A}}^k - \boldsymbol{\mathcal{R}} * \mathbf{Y}\big), \tag{22}$$

$$\boldsymbol{\mathcal{B}}^{k+1} = \xi_{\mu_2}\big(\boldsymbol{\mathcal{B}}^k - \boldsymbol{\mathcal{Q}} * \boldsymbol{\rho}_2^k + \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{B}}^k - \boldsymbol{\mathcal{R}} * \mathbf{Z}\big). \tag{23}$$

Similarly, the update of $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$ will take the form

$$\boldsymbol{\rho}_1^{k+1} = \boldsymbol{\rho}_1^k + \boldsymbol{\mathcal{T}}_1 * \boldsymbol{\mathcal{A}}^{k+1} - \mathbf{Y}, \tag{24}$$

$$\boldsymbol{\rho}_2^{k+1} = \boldsymbol{\rho}_2^k + \boldsymbol{\mathcal{T}}_2 * \boldsymbol{\mathcal{B}}^{k+1} - \mathbf{Z}. \tag{25}$$

Equations (22) - (25) can be easily translated into a neural network with convolutional layers parameterized by $\boldsymbol{\mathcal{Q}}$, $\boldsymbol{\mathcal{R}}$, $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{T}}_1$, $\boldsymbol{\mathcal{T}}_2$. The obtained CNN learns coupled sparse feature maps. Fig. 1 depicts one stage of the proposed architecture corresponding to a single iteration of the MM based algorithm.

By adding more stages to our design followed by a reconstruction layer, which computes the corresponding image using the learned feature maps, we obtain a multimodal deep CNN architecture for guided image super-resolution. Fig. 2 illustrates an instant of our multimodal CNN design performing three iterations of the proposed MM algorithm.

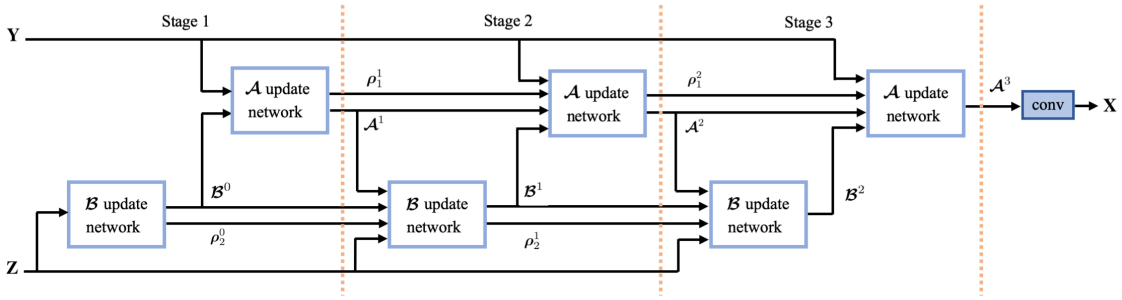


Fig. 2: The proposed multimodal network with three update stages and one reconstruction layer.

TABLE I: Performance comparison using multi-spectral test images for $\times 4$ upscaling factor.

Image	bicubic		GF [2]		SRFBN [12]		DJF [5]		CoISTA [18]		LMCSC [4]		Proposed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Chart toy	28.94	0.9424	34.09	0.9788	33.43	0.9838	37.86	0.9935	36.58	0.9914	40.31	0.9965	**41.13**	**0.9977**
Egyptian	36.57	0.9786	40.24	0.9796	40.04	0.9822	45.69	0.9922	45.91	0.9961	48.79	0.9981	**49.92**	**0.9991**
Feathers	30.80	0.9562	33.60	0.9748	35.53	0.9873	40.13	0.9939	39.62	0.9937	41.48	0.9962	**42.27**	**0.9974**
Glass tiles	26.65	0.9242	29.46	0.9593	29.53	0.9676	34.97	0.9915	33.99	0.9907	34.65	0.9939	**34.84**	**0.9948**
Jelly beans	27.81	0.9302	30.90	0.9658	32.97	0.9845	39.16	0.9885	38.92	0.9956	39.75	0.9966	**40.75**	**0.9980**
Oil Paintings	31.67	0.8943	35.03	0.9441	32.68	0.9182	37.76	0.9805	37.26	0.9690	39.14	0.9910	**39.84**	**0.9926**
Paints	29.29	0.9493	31.73	0.9702	36.06	0.9907	39.36	0.9944	38.40	0.9949	38.98	0.9966	**39.69**	**0.9978**
Average	30.25	0.9393	33.58	0.9675	34.32	0.9735	39.28	0.9906	38.67	0.9902	40.44	0.9955	**41.20**	**0.9968**

TABLE II: Performance comparison using multi-spectral test images for $\times 8$ upscaling factor.

Image	bicubic		GF [2]		SRFBN [12]		DJF [5]		CoISTA [18]		LMCSC [4]		Proposed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Chart toy	25.00	0.8048	29.77	0.9446	27.90	0.8736	32.89	0.9733	33.18	0.9768	34.35	0.9805	**35.47**	**0.9837**
Egyptian	33.12	0.9316	36.52	0.9627	35.50	0.9755	41.58	0.9850	43.46	0.9906	43.90	0.9966	**44.40**	**0.9969**
Feathers	25.59	0.8067	27.85	0.9183	30.14	0.9718	31.50	0.9396	32.04	0.9432	36.81	0.9875	**37.83**	**0.9901**
Glass tiles	22.56	0.7308	25.24	0.8641	23.72	0.9002	29.53	0.9685	27.96	0.9390	30.20	0.9724	**30.89**	**0.9797**
Jelly beans	23.04	0.7388	25.05	0.8813	25.80	0.9243	30.14	0.9503	30.69	0.9585	34.70	0.9888	**34.79**	**0.9890**
Oil Paintings	30.56	0.8234	33.82	0.9383	31.07	0.9258	35.12	0.9492	35.99	0.9482	36.27	0.9759	**36.98**	**0.9793**
Paints	26.40	0.8465	29.25	0.9451	28.03	0.9653	31.86	0.9553	33.05	0.9679	**35.06**	0.9910	34.96	**0.9927**
Average	26.61	0.8118	29.64	0.9221	28.88	0.9338	33.23	0.9602	33.77	0.9615	35.90	0.9847	**36.47**	**0.9873**

IV. EXPERIMENTS

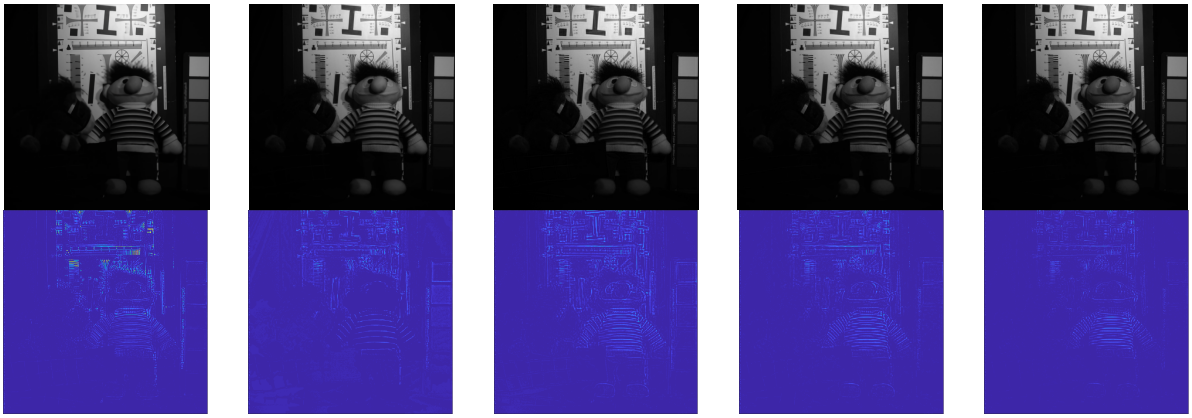
This section presents the implementation details for the proposed model and its performance evaluation. The model is employed for the super-resolution of multi-spectral data with the aid of HR RGB images. The experimental results include comparison with state-of-the-art methods.

We employ the Columbia multi-spectral database[1] for the experiments on guided multi-spectral image super-resolution. We create a training set consisting of $40,000$ image patches of size 64×64 extracted from the original images. We produce the LR version of the ground truth images from the target modality by performing blurring and downscaling operations. The guidance modality in our network only includes the luminance channel of the corresponding RGB image. We reserve seven pairs for testing. We apply SR at different scales, and train the network separately for every scale.

We realize a model with two stages corresponding to two iterations of the multimodal MM algorithm. We initialize the sparse representation $\boldsymbol{\mathcal{A}}$ of the target modality and the Lagrange parameters $\boldsymbol{\rho_1}$, $\boldsymbol{\rho_2}$ with zero. The convolutional layers contain 64 kernels of size 7×7. 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.

We evaluate the performance of our network against [12], a state-of-the-art deep model for single-modal image SR, and against several state-of-the-art multimodal image SR models, namely, guided image filtering (GF) [2], the deep joint image filtering (DJF) [5], the CoISTA network [18] and the learned multimodal convolutional sparse coding (LMCSC) model presented in our previous work [4]. Results for different upscaling factors are summarized in Tables I and II. As can be seen the proposed model notably outperforms the state of the art. Compared to the best competing method [4], the model brings

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



(f) SRFBN [12] (b) DJF [5] (a) CoISTA [18] (b) LMCSC [4] (c) Proposed

Fig. 3: Super-resolution of the multi-spectral image "chart toy" with upscaling factor $\times 4$ and the corresponding error maps.

average peak signal-to-noise-ratio (PSNR) improvements up to 0.76dB. The numerical results are corroborated by visual examples. Fig. 3 presents a super-resolved multi-spectral test image using five models together with their error maps.

V. CONCLUSION

We presented a coupled convolutional sparse coding scheme for multimodal image SR. The problem was addressed by a method of multipliers-based algorithm, which was unfolded into a multimodal CNN design. The proposed model consists of a few stages performing convolutional sparse coding operations and fusion of information coming from the two modalities. Our model was applied to multimodal image super-resolution of multi-spectral images with the aid of RGB images. Experimental results demonstrate the superior performance of the proposed multimodal CNN model against single-modal and multimodal designs.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Learned multimodal convolutional sparse coding for guided image super-resolution," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

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

[6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, pp. 2861–2873, 2010.

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

[8] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. S. Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

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

[10] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

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

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

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

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

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

[24] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multimodal deep unfolding for guided image super-resolution," *arXiv:2001.07575*, 2020.

[25] 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.