

Multimodal Image Super-resolution via Joint Sparse Representations induced by Coupled Dictionaries

Pingfan Song, *Student Member, IEEE*, Xin Deng, *Student Member, IEEE*,
 João F. C. Mota, *Member, IEEE*, Nikos Deligiannis, *Member, IEEE*,
 Pier Luigi Dragotti, *Fellow, IEEE*, and Miguel R. D. Rodrigues, *Senior Member, IEEE*

Abstract—Real-world data processing problems often involve various image modalities associated with a certain scene, including RGB images, infrared images or multi-spectral images. The fact that different image modalities often share certain attributes, such as certain edges, textures and other structure primitives, represents an opportunity to enhance various image processing tasks. This paper proposes a new approach to construct a high-resolution (HR) version of a low-resolution (LR) image given another HR image modality as reference, based on joint sparse representations induced by coupled dictionaries. Our approach, which captures the similarities and disparities between different image modalities in a learned sparse feature domain *in lieu* of the original image domain, consists of two phases. The coupled dictionary learning phase is used to learn a set of dictionaries that couple different image modalities in the sparse feature domain given a set of training data. In turn, the coupled super-resolution phase leverages such coupled dictionaries to construct a HR version of the LR target image given another related image modality. One of the merits of our sparsity-driven approach relates to the fact that it overcomes drawbacks such as the texture copying artifacts commonly resulting from inconsistency between the guidance and target images. Experiments on real multimodal images demonstrate that incorporating appropriate guidance information via joint sparse representation induced by coupled dictionary learning brings notable benefits in the super-resolution task with respect to the state-of-the-art. Of particular relevance, the proposed approach also demonstrates better robustness than competing deep-learning-based methods in the presence of noise.

Index Terms—Multimodal image super-resolution, coupled dictionary learning, joint sparse representation, side information

I. INTRODUCTION

Image super-resolution (SR) is an operation that involves the enhancement of pixel-based image resolution, while minimizing visual artifacts. However, the construction of a high-resolution (HR) version of a low-resolution (LR) image requires inferring the values of missing pixels, making image

This work is supported by China Scholarship Council (CSC), UCL Overseas Research Scholarship (UCL-ORS), the VUB-UGent-UCL-Duke International Joint Research Group grant, and by EPSRC grant EP/K033166/1.

Pingfan Song and Miguel R. D. Rodrigues are with the Department of Electronic & Electrical Engineering, University College London, London WC1E 6BT, UK. (e-mail: pingfan.song.14@ucl.ac.uk, m.rodrigues@ucl.ac.uk)

Xin Deng and Pier Luigi Dragotti are with the Department of Electronic & Electrical Engineering, Imperial College London, London SW7-2AZ, UK. (e-mail: x.deng16@imperial.ac.uk, p.dragotti@imperial.ac.uk)

João F. C. Mota is with the Department of School of Engineering & Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK. (email: j.mota@hw.ac.uk)

N. Deligiannis is with the Department of Electronics and Informatics, Vrije Universiteit Brussel, B-1050 Brussels, Belgium, and with imec, Kapeldreef 75, B-3001 Leuven, Belgium. (e-mail: ndeliglia@etrvub.be)

SR a severely ill-posed problem. Various image models and approaches have been proposed to regularize this ill-posed problem via employing some prior knowledge, including natural priors [1]–[4], local and non-local similarity [5], [6], sparse representation over fixed or learned dictionaries [7]–[13], and sophisticated features from deep learning [14]–[17]. These typical super-resolution approaches focus only on single modality images without exploiting the availability of other modalities as guidance.

However, in many practical application scenarios, a certain scene is often imaged using different sensors yielding different image modalities. For example, in remote sensing it is typical to have various image modalities of earth observations, such as a panchromatic band version, a multi-spectral bands version, and an infrared (IR) band version [18], [19]. In order to balance cost, bandwidth and complexity, these multimodal images are usually acquired with different resolutions [18]. These scenarios call for approaches that can capitalize on the availability of multiple image modalities of the same scene – which typically share textures, edges, corners, boundaries, or other salient features – in order to super-resolve the LR images with the aid of the HR images of a different modality.

Therefore, a variety of joint super-resolution/upsampling approaches have been proposed to leverage the availability of additional *guidance images*, also referred to as *side information* [20], [21], to aid the super-resolution of target LR modalities [22]–[27]. The basic idea behind these methods is that the structural details of the guidance image can be transferred to the target image. However, these methods tend to introduce notable texture-copying artifacts, i.e. erroneous structure details that are not originally present in the target image because such methods typically fail to distinguish similarities and disparities between the different image modalities.

The motivation of this work is to introduce a new image SR approach, based on joint sparse representations induced by coupled dictionaries, that has the ability to take into account both similarities and disparities between target and guidance images in order to deliver superior SR performance.

Proposed Scheme. The proposed scheme is based on three elements: (1) a data model; (2) a coupled dictionary learning algorithm; and (3) a coupled image super-resolution algorithm.

- **Data Model:** This is a patch-based model that relies on the use of coupled dictionaries to jointly sparsely represent a pair of patches from the different image modalities. Of particular relevance is the ability to represent the similarities and disparities between the different image

modalities in this sparse feature domain in *lieu* of the original image domain, which leads to a higher super-resolution accuracy.

- *Coupled Dictionary Learning*: This algorithm learns the data model – including a set of coupled dictionaries along with the joint sparse representations of the different image modalities – from a set of training images.
- *Coupled Image Super-Resolution*: This algorithm uses the learned coupled dictionaries to perform joint sparse coding for the target/guidance image pair. The resulting joint sparse representations are then used to estimate a HR version of the target image from its LR version.

In comparison with state-of-the-art approaches [22]–[26], our approach can better model the common and distinct features of the different data modalities. This capability makes our approach more robust to inconsistencies between the guidance and the target images, as both the target LR image and the guidance image are taken into account during the estimation of the target HR image, instead of unilaterally transferring the structure details from the HR guidance image. In addition, our approach is also more robust to mismatches between training data and testing data (e.g. due to the presence of noise) in comparison to deep-learning-based approaches [25].

Contributions. Our contributions are as follows:

- We devise a data model for multimodal signals that captures the similarities and disparities between different modalities using joint sparse representations induced by coupled dictionaries. Compared with our previous work [28], the present model is more general, because it does not require the matrix that models the conversion of a HR version of the image to the LR counterpart to be known.
- We also propose a learning algorithm to learn the coupled dictionaries from different data modalities. Again, compared with [28], in the learning stage, the proposed algorithm does not require the knowledge of the matrix that converts a HR image to a LR version.
- We also propose a multimodal image super-resolution algorithm that enhances the resolution of the target LR image with the aid of another guidance HR image modality.
- Finally, extensive experiments are conducted both on a variety of multimodal images. The results demonstrate that our proposed approach leads to better super-resolution performance than state-of-the-art approaches in a range of scenarios.

Organization. The remainder of this paper is organized as follows. We review related work in Section II, including single and joint image SR, as well as other multimodal image processing works. We then propose our multimodal image super-resolution framework, including the data model, the coupled dictionary learning algorithm, and the multimodal image super-resolution algorithm in Section III. Section IV is devoted to various simulation and practical experiments which demonstrate that our approach can lead to significant gains over the state of the art. We summarize the main contributions of the paper in Section V.

II. RELATED WORK

There are various image super-resolution approaches in the literature. Single image super-resolution approaches do not leverage other guidance images, whereas joint image super-resolution approaches explicitly leverage the availability of other image modalities.

A. Single image SR

In general, conventional single image SR approaches can be categorized into three classes: (1) interpolation-based, (2) reconstruction-based and (3) learning-based SR approaches.

Interpolation-based SR approaches. Advanced interpolation approaches exploit natural image priors, such as edges [1], image smoothness [2], gradient profile [3] and other geometric regularity of image structures [4]. These methods are simple and fast, but tend to overly smooth image edges and generate ringing and jagged artifacts.

Reconstruction-based SR approaches. Reconstruction-based SR approaches, also referred to as model-based SR methods, attempt to regularize the highly under-determined image SR inverse problem by exploiting various image priors, including self-similarity of images patches [5], sparsity in the wavelet domain [7], analysis operator [29], and other fused versions [6]. Recent work [12] proposes a piecewise smooth image model and makes use of the finite rate of innovation (FRI) theory to reconstruct HR target images. These reconstruction-based methods usually offer better performance than interpolation-based methods.

Learning-based SR approaches. These SR approaches typically consist of two phases: (1) a learning phase where one learns certain image priors from training images and (2) a testing phase where one obtains the HR image from the LR version with the aid of the prior knowledge.

In particular, patch-wise learning-based approaches leverage learned mappings or co-occurrence priors between LR and HR training image patches to predict the fine details in the testing target HR images according to their corresponding LR versions [8]–[11], [13], [30]–[33]. For example, motivated by Compressive Sensing [34], [35], Yang *et al.* [8], [9], [31] propose a sparse-coding based image SR strategy, which is improved further by Zeyde, *et al.* [10]. The key idea is a sparse representation invariance assumption which states that HR/LR image pairs share the same sparse coefficients with respect to a pair of HR and LR dictionaries. Along similar lines, Timofte *et al.* [11], [13] propose a strategy, referred to as anchored neighbourhood regression, that combines the advantage of neighbor embedding and dictionary learning. In order to achieve better flexibility and stability of signal recovery, semi-coupled dictionary learning [32] and coupled dictionary learning [33] are proposed to relax the sparse representation invariance assumption to the same support assumption, allowing more flexible mappings. Note that, even though the terminology related to "coupled dictionary learning" also appears in these works [9], [32], [33], their approaches focus only on coupling LR and HR images of the same modality, and do not take advantage of other image modalities. In addition, their assumptions, models and algorithms are also different from ours.

Inspired by sparse-coding-based SR methods, Dong *et al.* [14] propose a single image super-resolution convolutional neural network (SRCNN) consisting of a patch extraction and representation layer, a non-linear mapping layer and a reconstruction layer. A faster and deeper version FSRCNN was proposed in [15], where the previous interpolation operation is removed and a deconvolution layer is introduced at the end of the network to perform upsampling. Kim *et al.* [16] propose a very deep SR network (VDSR) which exploits residual-learning for fast converging and multi-scale training datasets for handling multiple scale factors. Different from the above CNN-based SR approaches, [17] proposes a deeply-recursive convolutional network (DRCN) with recursive-supervision and skip-connection to ease the training.

B. Joint image SR

Compared with single image SR, joint image SR attempts to leverage an additional guidance image to aid the SR process for the target image, by transferring structural information of the guidance image to the target image.

The bilateral filter [36] is a widely used translation-variant edge-preserving filter that outputs a pixel as a weighted average of neighboring pixels. The weights are computed by a spatial filter kernel and a range filter kernel evaluated on the data values themselves. It smoothes the image while preserving edges. The joint bilateral upsampling [22] generalizes the bilateral filter by computing the weights with respect to another guidance image rather than the input image. In particular, it applies the range filter kernel to a HR guidance image, expecting to incorporate the high frequencies of the guidance image into the LR target image. However, it has been noticed that joint bilateral image filtering may introduce gradient reversal artifacts as it does not preserve gradient information [23]. Later, guided image filtering [23] was proposed to overcome this limitation. Directly transferring guidance gradients can also result in notable appearance change [26]. To address this problem, [26] proposes a framework that optimizes a novel scale map to capture the nature of structure discrepancy between images. However, as the construction of these filters considers unilaterally the static guidance image, [26] suffers from the inconsistency of the local structures in the guidance and target images, and may therefore transfer incorrect structure details to the target images. The study in [24] proposes robust guided image filtering, referred to as static/dynamic (SD) filtering, which jointly leverages static guidance image and dynamic target image to iteratively refine the target image. These techniques use hand-crafted objective functions that may not reflect natural image priors well.

Recent work [25] proposes a Convolutional Neural Network (CNN) based joint image filtering approach. This approach considers the structures of both input and guidance images, but requires numerous labelled images and intensive computing resources to train the deep model for each task.

Our joint image SR based on coupled dictionary learning falls into the learning-based category. Therefore, the priors used in our approach are learned from a training dataset rather than being hand-crafted and thus adapt to the target modality and guidance modality.

C. Other multimodal image processing approaches based on sparse representations induced by a set of dictionaries

A number of multimodal image processing approaches based on sparse representations induced by a set of dictionaries have also been proposed in the literature [32], [37]–[43]. However, these approaches differ from our proposed approach in a number of ways. For example, semi-couple dictionary learning [32], supervised coupled dictionary learning [37], semi-supervised coupled dictionary learning [38], and semi-coupled low-rank discriminant dictionary learning [39] assume the existence of a function that maps the sparse representation of one modality to the sparse representation of another modality. In contrast, our approach does not constrain the model to require the existence of such a mapping function; instead, both similarities and disparities between different modalities are considered under the sparse representation invariance assumption. In turn, Dao *et al.* [40] propose a joint/collaborative sparse representation framework for multi-sensor classification. However, the dictionaries used in their work are directly constructed from training data samples and involve no dictionary learning. In comparison, the dictionaries in our work are learned from training data. Moreover, Bahrampour *et al.* [41] propose a multimodal task-driven dictionary learning algorithm under the group sparsity prior to enforce collaborations among multiple homogeneous/heterogeneous sources of information. One common feature of these works is that the sparse representations for different modalities are required to share the same support, usually induced by group sparsity, and their values are related by a mapping function.

In comparison, our model takes into account both the similarity and the discrepancy of different modalities via considering their common and unique sparse representations. This makes our approach more robust to inconsistencies between the guidance and the target images, as both of them are considered during the estimation of the target HR image, instead of unilaterally transferring the structure details from the guidance image. Our data model used in our multimodal image SR approach is inspired from the data model proposed in [42], [43] used for multimodal image separation. However, the generalization of the approach from multimodal image separation to multimodal image SR entails a number of innovations including: (1) unique dictionaries are introduced for the side information because we consider that the side information also contains its own unique features; (2) both our coupled dictionary learning and coupled SR algorithms are different from [42], [43]. Overall, practical experiments demonstrate that the proposed multimodal image SR approach outperforms the state-of-the-art in various scenarios.

III. MULTIMODAL IMAGE SR VIA JOINT SPARSE REPRESENTATIONS INDUCED BY COUPLED DICTIONARIES

We now introduce our SR approach. In particular, we describe the data model that couples different image modalities and also the joint image SR framework that encompasses both a coupled dictionary learning phase and a coupled super-resolution phase, see also Figure 1).

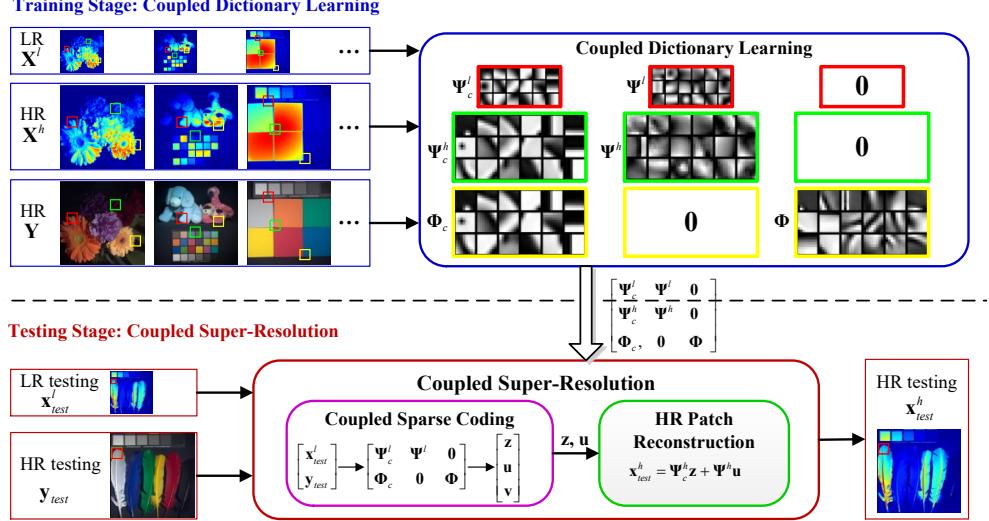


Figure 1: Proposed multimodal image super-resolution approach encompassing both a training stage and a testing stage. \mathbf{X} (or \mathbf{x}) and \mathbf{Y} (or \mathbf{y}) represent the target and guidance modalities, respectively.

A. Multimodal Data Model

Basic Data Model. It is commonly observed that images of different modalities contain similarities as well as disparities. These characteristics can be effectively modelled in a sparse feature space so that different modalities can be related together via their sparse representations with respect to a group of coupled dictionaries. We first introduce a basic data model that captures the relationships – including similarities and disparities – between two different image modalities. In particular, we propose to use joint sparse representations to express a pair of registered, vectorized image patches $\mathbf{x} \in \mathbb{R}^{N_x}$ and $\mathbf{y} \in \mathbb{R}^{N_y}$ associated with different modalities as follows:

$$\begin{aligned} \mathbf{x} &= \Psi_c \mathbf{z} + \Psi \mathbf{u}, \\ \mathbf{y} &= \Phi_c \mathbf{z} + \Phi \mathbf{v}, \end{aligned} \quad (1)$$

where $\mathbf{z} \in \mathbb{R}^{K_c}$ is a sparse representation that is common to both modalities, $\mathbf{u} \in \mathbb{R}^{K_u}$ is a sparse representation specific to modality \mathbf{x} , while $\mathbf{v} \in \mathbb{R}^{K_v}$ is a sparse representation specific to modality \mathbf{y} . In turn, $\Psi_c \in \mathbb{R}^{N_x \times K_c}$ and $\Phi_c \in \mathbb{R}^{N_y \times K_c}$ are a pair of dictionaries associated with the common sparse representation \mathbf{z} , whereas $\Psi \in \mathbb{R}^{N_x \times K_u}$ and $\Phi \in \mathbb{R}^{N_y \times K_v}$ are dictionaries associated with the specific sparse representations \mathbf{u} and \mathbf{v} , respectively. (For simplicity, we take $N = N_x = N_y$, $K = K_c = K_u = K_v$ hereafter.)

SR Data Model. We now transform the basic data model in (1) into the SR data model that underlies our proposed super-resolution process. This model is based on two main assumptions:

1. First, we assume – as in (1) – that similarities and disparities between the LR and HR versions of the patches of the different image modalities can be captured using sparse representations.
2. Second, we also assume – as in [8], [10], [31] – that the LR and HR versions of a patch of a certain image modality share the same sparse representation, albeit not the same dictionary.

In particular, we express the LR image patch $\mathbf{x}^l \in \mathbb{R}^M$ and HR image patch $\mathbf{x}^h \in \mathbb{R}^N$ of a certain image modality, and another HR registered patch of another corresponding image modality $\mathbf{y} \in \mathbb{R}^N$ as follows:¹

$$\mathbf{x}^h = \Psi_c^h \mathbf{z} + \Psi^h \mathbf{u}, \quad (2)$$

$$\mathbf{x}^l = \Psi_c^l \mathbf{z} + \Psi^l \mathbf{u}, \quad (3)$$

$$\mathbf{y} = \Phi_c \mathbf{z} + \Phi \mathbf{v}, \quad (4)$$

where, as in the basic data model (1), $\mathbf{z} \in \mathbb{R}^K$ is the common sparse representation shared by both modalities, $\mathbf{u} \in \mathbb{R}^K$ is the unique sparse representation specific to modality \mathbf{x} while $\mathbf{v} \in \mathbb{R}^K$ is the unique sparse representation specific to modality \mathbf{y} . In turn, $\Psi_c^h \in \mathbb{R}^{N \times K}$, $\Psi_c^l \in \mathbb{R}^{M \times K}$ and $\Phi_c \in \mathbb{R}^{N \times K}$ are the dictionaries associated with the common sparse representation \mathbf{z} , whereas $\Psi^h \in \mathbb{R}^{N \times K}$, $\Psi^l \in \mathbb{R}^{M \times K}$ and $\Phi \in \mathbb{R}^{N \times K}$ are dictionaries associated with the specific sparse representations \mathbf{u} and \mathbf{v} , respectively. Note that the sparse vectors \mathbf{z} and \mathbf{u} capture the relationship between the LR and HR patches of the same modality in (2) and (3). Moreover, the common sparse vector \mathbf{z} connects the various patches of the two different modalities in (2) – (4). The disparities between modalities \mathbf{x} and \mathbf{y} are distinguished by the sparse vectors \mathbf{u} and \mathbf{v} . Overall, this data model allows each pair of patches to be non-linearly transformed to a sparse domain with respect to a group of coupled dictionaries in order to obtain sparse representations that characterize the similarities and disparities between different modalities. Note also that our data model reduces to the data model in [8]–[10] – applicable to single modality image super-resolution – provided that the side information \mathbf{y} is neglected.

¹Our model assumes identical common sparse representations so that each pair of common atoms is adjusted automatically to satisfy this assumption. In addition, we also take into account the discrepancy of different modalities via considering their unique sparse representations. This differs from the models used in [32], [37]–[43], some of which assume that the sparse representations for different modalities share the same support and some assume that they share identical sparse representations without consideration to the discrepancy.

By capitalizing on this model, we propose in the sequel a novel joint image SR scheme that consists of two stages: (1) a training stage referred to as coupled dictionary learning (CDL) and (2) a testing stage referred to as coupled image super-resolution (CSR) (see Figure 1). In the training stage, we learn the dictionaries in (2) - (4) from a set of training image patches to couple different data modalities together. Then, in the testing stage, we use the learned dictionaries to find the representations of the LR testing patch and corresponding HR guidance patch, according to (3) and (4). These sparse representations are then used to reconstruct the desired HR target image patch via (2).

B. Coupled Dictionary Learning (CDL)

We assume that we have access to T registered patches of LR, HR and guidance images for learning our data model in (2) - (4). In particular, let \mathbf{x}_i^l , \mathbf{x}_i^h and \mathbf{y}_i ($i = 1 \dots T$) denote the registered patches corresponding to the LR, HR, and the guidance training image patches, and let \mathbf{z}_i , \mathbf{u}_i and \mathbf{v}_i ($i = 1 \dots T$) denote their sparse representations. Our coupled dictionary learning problem can now be posed as follows:

$$\begin{aligned} & \underset{\{\Psi_c^l, \Psi^l, \Psi_c^h, \Phi_c, \Phi\}}{\text{minimize}} \quad \left\| \begin{bmatrix} \mathbf{X}^l \\ \mathbf{X}^h \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \Psi_c^l & \Psi^l & 0 \\ \Psi_c^h & \Psi^h & 0 \\ \Phi_c & 0 & \Phi \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \\ \mathbf{V} \end{bmatrix} \right\|_F^2 \\ & \text{subject to} \quad \|\mathbf{z}_i\|_0 + \|\mathbf{u}_i\|_0 + \|\mathbf{v}_i\|_0 \leq s, \quad \forall i, \end{aligned} \quad (5)$$

where $\mathbf{X}^l = [\mathbf{x}_1^l, \dots, \mathbf{x}_T^l] \in \mathbb{R}^{M \times T}$, $\mathbf{X}^h = [\mathbf{x}_1^h, \dots, \mathbf{x}_T^h] \in \mathbb{R}^{N \times T}$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T] \in \mathbb{R}^{N \times T}$, $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_T] \in \mathbb{R}^{K \times T}$, $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_T] \in \mathbb{R}^{K \times T}$ and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_T] \in \mathbb{R}^{K \times T}$, and $\|\cdot\|_F$ and $\|\cdot\|_0$ denote the Frobenius norm and ℓ_0 pseudo-norm, respectively.

Note that – akin to other dictionary learning formulations [44] – the objective in the optimization problem (5) encourages the data representation to approximate the data, and the constraint in (5) encourages the data representation to be sparse (i.e. the overall sparsity of the data representations is constrained to be less than or equal to s)².

We address the coupled dictionary learning problem (5) in two steps: LR Dictionary learning and HR Dictionary learning. In the first step (LR Dictionary learning), the algorithm uses LR patches \mathbf{X}^l and side information \mathbf{Y} to learn the two pairs of dictionaries $[\Psi_c^l, \Psi^l]$ and $[\Phi_c, \Phi]$ and the sparse codes \mathbf{Z} , \mathbf{U} , \mathbf{V} , via solving a non-convex optimization problem. In the second step (HR Dictionary learning), the algorithm uses HR patches \mathbf{X}^h and the sparse codes \mathbf{U} , \mathbf{V} to learn the HR dictionaries $[\Psi_c^h, \Psi^h]$.³ Algorithm 1 shows how we adapt K-SVD [46] accordingly.

²Note that, we could also use alternative sparsity constraints, such as (a) $\|\mathbf{z}_i\|_0 + \|\mathbf{u}_i\|_0 \leq s_x$, $\|\mathbf{z}_i\|_0 + \|\mathbf{v}_i\|_0 \leq s_y$, (b) $\|\mathbf{z}_i\|_0 \leq s_z$, $\|\mathbf{u}_i\|_0 \leq s_u$, $\|\mathbf{v}_i\|_0 \leq s_v$. Empirical studies suggest that these constraints lead to similar performance. We prefer the constraint in (5) since it makes the formulation concise, with fewer parameters for tuning.

³The motivation of this two-step training strategy is that the sparse codes \mathbf{Z} and \mathbf{U} should be obtained only from \mathbf{X}^l and \mathbf{Y} in both training and testing stages without involving \mathbf{X}^h , since the HR target patches \mathbf{X}^h are available only in the training stage and not in testing stage. Similar strategies are also adopted by other works [10] and the empirical results suggest better performance.

1) Step 1 – LR Dictionary learning: In the first step, we learn the dictionary pairs $[\Psi_c^l, \Psi^l]$, $[\Phi_c, \Phi]$ and the sparse codes \mathbf{Z} , \mathbf{U} , \mathbf{V} from \mathbf{X}^l and \mathbf{Y} by solving the following optimization problem:

$$\begin{aligned} & \underset{\{\Psi_c^l, \Psi^l, \Phi_c, \Phi\}}{\text{minimize}} \quad \left\| \begin{bmatrix} \mathbf{X}^l \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \Psi_c^l & \Psi^l & 0 \\ \Phi_c & 0 & \Phi \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \\ \mathbf{V} \end{bmatrix} \right\|_F^2 \\ & \text{subject to} \quad \|\mathbf{z}_i\|_0 + \|\mathbf{u}_i\|_0 + \|\mathbf{v}_i\|_0 \leq s, \quad \forall i. \end{aligned} \quad (6)$$

In order to handle this non-convex optimization problem, we adopt an alternating optimization approach that performs sparse coding and dictionary update alternatively.

During the sparse coding stage, we first fix the global dictionaries and obtain the sparse representations by solving:

$$\begin{aligned} & \underset{\mathbf{Z}, \mathbf{U}, \mathbf{V}}{\text{minimize}} \quad \left\| \begin{bmatrix} \mathbf{X}^l \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \Psi_c^l & \Psi^l & 0 \\ \Phi_c & 0 & \Phi \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \\ \mathbf{V} \end{bmatrix} \right\|_F^2 \\ & \text{s.t.} \quad \|\mathbf{z}_i\|_0 + \|\mathbf{u}_i\|_0 + \|\mathbf{v}_i\|_0 \leq s, \quad \forall i. \end{aligned} \quad (7)$$

This problem – which we call global sparse coding because it updates all the sparse representations \mathbf{Z} , \mathbf{U} and \mathbf{V} – is solved using the orthogonal matching pursuit (OMP) algorithm [45].⁴

During the dictionary updating stage, we fix the sparse codes and update the global dictionaries via solving:

$$\underset{\Psi_c^l, \Psi^l, \Phi_c, \Phi}{\text{minimize}} \quad \left\| \begin{bmatrix} \mathbf{X}^l \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \Psi_c^l & \Psi^l & 0 \\ \Phi_c & 0 & \Phi \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \\ \mathbf{V} \end{bmatrix} \right\|_F^2. \quad (8)$$

To this end, we adapt the K-SVD [44] algorithm for our coupled dictionary learning case. The key idea is to update common dictionaries simultaneously while updating unique dictionaries individually.⁵ Specifically, we further decompose Problem (6) into the following convex sub-problems (9) - (11), so that we can sequentially learn the common dictionaries and the unique dictionaries. That is, we fix the unique dictionaries Ψ^l , Φ and only update the common dictionaries Ψ_c^l and Φ_c by solving

$$\underset{\Psi_c^l, \Phi_c}{\text{minimize}} \quad \left\| \begin{bmatrix} \mathbf{X}^l - \Psi^l \mathbf{U} \\ \mathbf{Y} - \Phi \mathbf{V} \end{bmatrix} - \begin{bmatrix} \Psi_c^l \\ \Phi_c \end{bmatrix} \mathbf{Z} \right\|_F^2. \quad (9)$$

The algorithm alternates between global sparse coding (7) and local common dictionary update (9) for a few iterations until the procedure converges. Next, we fix the already learned common dictionaries and train the unique dictionaries by alternating between global sparse coding (7) and following two unique dictionary update operations:

$$\underset{\Psi^l}{\text{minimize}} \quad \|(\mathbf{X}^l - \Psi_c^l \mathbf{Z}) - \Psi^l \mathbf{U}\|_F^2. \quad (10)$$

$$\underset{\Phi}{\text{minimize}} \quad \|(\mathbf{Y} - \Phi_c \mathbf{Z}) - \Phi \mathbf{V}\|_F^2. \quad (11)$$

⁴An additional error threshold parameter ϵ is used to deal with noisy images. This parameter defines whether or not one should stop the OMP loop depending on the residual of the objective. See Algorithm 1.

⁵Owing to the SVD operation in the dictionary update, atoms from the common dictionary pair $[\Psi_c^l, \Phi_c]$ and the unique dictionaries Ψ^l and Φ have unit ℓ_2 norm automatically.

Algorithm 1 Coupled Dictionary Learning

Input: Training data matrices \mathbf{X}^l , \mathbf{X}^h and \mathbf{Y} .
Output: Dictionary pairs $[\Psi_c^l, \Psi^l]$, $[\Psi_c^h, \Psi^h]$ and $[\Phi_c, \Phi]$.
Initialization: Initialize dictionary atoms with randomly selected patches. Set the training iterations $OutIter$ and $InIter$, sparsity constraint s and residual constraint ϵ .
Optimization:

- 1: **Step 1 – LR Dictionary learning:**
- 2: **for** $p = 1$ to $OutIter$ **do**
- 3: **for** $q = 1$ to $InIter$ **do**
- 4: **Global Sparse Coding.** Fix all the dictionaries, then solve (7) to update sparse representations \mathbf{Z} , \mathbf{U} and \mathbf{V} by performing OMP on each training example.
- 5: Initialize the active set $\Gamma = \emptyset$ and $[\mathbf{z}_i^T; \mathbf{u}_i^T; \mathbf{v}_i^T] \leftarrow \mathbf{0}$.
- 6: **while** $|\Gamma| < s_c$ or residual $> \epsilon$ **do**
- 7: select a new coordinate \hat{k} that leads to the smallest residual and, then update the active set and the sparse representations:

$$(\hat{k}, \hat{\alpha}) \in \arg \min_{k \in \Gamma^c, \alpha \in \mathbb{R}^{|\Gamma|+1}} \left\| \begin{bmatrix} \mathbf{x}_i^l \\ \mathbf{y}_i \\ \mathbf{\Phi}_c^l \mathbf{\Psi}^l \mathbf{0} \end{bmatrix}_{\Gamma \cup \{k\}} - \begin{bmatrix} \mathbf{\Psi}_c^l \mathbf{\Psi}^l \mathbf{0} \\ \mathbf{\Phi}_c^l \mathbf{0} \mathbf{\Phi}_c^l \end{bmatrix}_{\Gamma \cup \{k\}} \alpha \right\|_2^2$$

$$\Gamma \leftarrow \Gamma \cup \{\hat{k}\}; [\mathbf{z}_i^T; \mathbf{u}_i^T; \mathbf{v}_i^T]_{\Gamma} \leftarrow \hat{\alpha}; [\mathbf{z}_i^T; \mathbf{u}_i^T; \mathbf{v}_i^T]_{\Gamma^c} \leftarrow \mathbf{0}$$

- 8: **end while**
- 9: **Local Common Dictionary Update.** Fix Ψ^l , Φ , and only update Ψ_c^l and Φ_c by solving (9). Specifically, for each atom pair $\begin{bmatrix} \psi_{ck}^l \\ \phi_{ck}^l \end{bmatrix}$ of $\begin{bmatrix} \Psi_c^l \\ \Phi_c \end{bmatrix}$, denote by \mathbf{z}^k the k -th row vector in \mathbf{Z} , and $\Omega_k = \{i | 1 \leq i \leq T, \mathbf{z}^k(i) \neq 0\}$ the index set of those training samples

2) *Step 2 – HR Dictionary learning:* In the second step, once the dictionary pairs $[\Psi_c^l, \Psi^l]$ and $[\Phi_c, \Phi]$ are learned from \mathbf{X}^l and \mathbf{Y} , we construct the HR dictionaries $[\Psi_c^h, \Psi^h]$ based on \mathbf{X}^h and sparse codes \mathbf{Z} and \mathbf{U} by solving the optimization problem:

$$\min_{\Psi_c^h, \Psi^h} \|\mathbf{X}^h - \Psi_c^h \mathbf{Z} - \Psi^h \mathbf{U}\|_F^2 + \lambda \|\begin{bmatrix} \Psi_c^h & \Psi^h \end{bmatrix}\|_F^2 \quad (12)$$

where the second term serves as a regularizer that makes the solution more stable⁶. This optimization problem – which exploits the conventional sparse representation invariance assumption that HR image patches \mathbf{X}^h share the same sparse codes with the corresponding LR version \mathbf{X}^l – admits the closed form solution

$$[\Psi_c^h \quad \Psi^h] = \mathbf{X}^h \Gamma^T (\Gamma \Gamma^T + \lambda \mathbf{I})^{-1}, \text{ where, } \Gamma = \begin{bmatrix} \mathbf{Z} \\ \mathbf{U} \end{bmatrix} \quad (13)$$

Similar to conventional dictionary learning, our CDL algorithm cannot guarantee the convergence to a global optimum

⁶In order to guarantee the fidelity of the sparse approximation to the HR training datasets, the atoms in the HR dictionaries are not constrained to be unit ℓ_2 norm, as in [10] and [46]. However, when there are zeros or near-zeros rows in the sparse codes \mathbf{Z} or \mathbf{U} , the matrix inverse operation during the computation of the closed form solution will give extremely large value for corresponding atoms. Therefore, in order to make the solution more stable, a Frobenius norm is added to regularize Problem (12).

that use k -th atom pair. Then, compute the representation residual

$$\mathbf{E}_k = \left(\begin{bmatrix} \mathbf{X}^l - \Psi^l \mathbf{U} \\ \mathbf{Y} - \Phi \mathbf{V} \end{bmatrix} - \begin{bmatrix} \Psi_c^l \\ \Phi_c \end{bmatrix} \mathbf{Z} + \begin{bmatrix} \psi_{ck}^l \\ \phi_{ck}^l \end{bmatrix} \mathbf{z}^k \right)_{(:, \Omega_k)}$$

Apply SVD on $\mathbf{E}_k = \mathbf{P} \mathbf{\Sigma} \mathbf{Q}^T$ and choose the first column of \mathbf{P} as the updated atom pair $\begin{bmatrix} \psi_{ck}^l \\ \phi_{ck}^l \end{bmatrix}$.

- 10: **end for**
- 11: **for** $q = 1$ to $InIter$ **do**
- 12: **Global Sparse Coding.** The same as step 4.
- 13: **Local Unique Dictionary Update.** Fix Ψ_c^l , Φ_c , and only update Ψ^l and Φ by solving (10) and (11). For each atom ψ_k^l of Ψ^l , denote by \mathbf{u}^k the k -th row vector in \mathbf{U} , and $\Omega_k = \{i | 1 \leq i \leq T, \mathbf{u}^k(i) \neq 0\}$. Then, compute the representation residual

$$\mathbf{E}_k = ([\mathbf{X}^l - \Psi_c^l \mathbf{Z}] - \Psi^l \mathbf{U} + \psi_k^l \mathbf{u}^k)_{(:, \Omega_k)}$$

Apply SVD on $\mathbf{E}_k = \mathbf{P} \mathbf{\Sigma} \mathbf{Q}^T$ and choose the first column of \mathbf{P} as the updated atom ψ_k^l . Each atom ϕ_k of Φ is updated with $\Omega_k = \{i | 1 \leq i \leq T, \mathbf{v}^k(i) \neq 0\}$ and $\mathbf{E}_k = ([\mathbf{Y} - \Phi_c \mathbf{Z}] - \Phi \mathbf{V} + \phi_k \mathbf{v}^k)_{(:, \Omega_k)}$ in a similar manner.

- 14: **end for**
- 15: **end for**
- 16: **Step 2 – HR Dictionary learning:**
- 17: Construct $[\Psi_c^h, \Psi^h]$ as in (13).
- 18: Return dictionaries.

due to the non-convexity nature of Problem (5). However, CDL is convex with respect to the dictionaries when the sparse codes are fixed or vice versa. This property ensures that dictionary algorithms usually converge to a local optimum that leads to good SR performance. This is also confirmed by experiments on both real and synthetic data, presented in Section IV.

C. Coupled Super Resolution (CSR)

Given the learned coupled dictionaries associated with the model in (2) - (4), we now assume that we have access to a LR testing image and a corresponding registered HR guidance image as side information. We extract (overlapping) image patch pairs from these two modalities. In particular, let $\mathbf{x}_{test}^l \in \mathbb{R}^M$ denote a LR testing image patch and let $\mathbf{y}_{test}^h \in \mathbb{R}^N$ denote the corresponding HR guidance image patch. We can now pose a coupled super-resolution problem that involves two steps.

1) *Step 1 – Coupled Sparse Coding:* First, we solve the optimization problem

$$\begin{aligned} \min_{\mathbf{z}, \mathbf{u}, \mathbf{v}} \quad & \left\| \begin{bmatrix} \mathbf{x}_{test}^l \\ \mathbf{y}_{test}^h \end{bmatrix} - \begin{bmatrix} \Psi_c^l & \Psi^l & \mathbf{0} \\ \Phi_c & \mathbf{0} & \Phi \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{u} \\ \mathbf{v} \end{bmatrix} \right\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{z}\|_0 + \|\mathbf{u}\|_0 + \|\mathbf{v}\|_0 \leq s, \end{aligned} \quad (14)$$

Algorithm 2 Coupled Super-resolution

Input: The testing patch \mathbf{x}_{test}^l and side information \mathbf{y}_{test} .
Learned dictionaries $[\Psi_c^l, \Psi^l]$, $[\Psi_c^h, \Psi^h]$ and $[\Phi_c, \Phi]$.

Output: High resolution estimation \mathbf{x}_{test}^h .

Operations:

1: **Step 1 – Coupled Sparse Coding:**

Use off-the-shelf sparse coding algorithms to solve the problem (14) to obtain the sparse codes \mathbf{z} , \mathbf{u} and \mathbf{v} .

2: **Step 2 – HR Patch Reconstruction:**

Reconstruct the HR patch \mathbf{x}_{test}^h as in (15).

where the ℓ_2 norm promotes the fidelity of sparse representations to the signals and the ℓ_0 pseudo-norm promotes sparsity for the sparse codes. Some off-the-shelf algorithms – such as orthogonal matching pursuit (OMP) algorithm [45] and iterative hard-thresholding algorithm [47] – can be applied to approximate the solution to (14). Compared with conventional sparse coding problems that involves only LR image patch \mathbf{x}^l , our formulations (14) also integrates the side information \mathbf{y}_{test} into the sparse coding task. Since the increase in the amount of available information is akin to the increase of the number of measurements in a Compressive Sensing scenario [20], [21], one can expect to obtain a more accurate estimate of the sparse codes.

2) *Step 2 – HR Patch Reconstruction:* Finally, we can obtain an estimate of the HR patch of the target image \mathbf{x}_{test}^h from the HR dictionaries $[\Psi_c^h, \Psi^h]$ and sparse codes \mathbf{z} and \mathbf{u} as follows

$$\mathbf{x}_{test}^h = \Psi_c^h \mathbf{z} + \Psi^h \mathbf{u}. \quad (15)$$

Once all the HR patches are recovered, they are integrated into a whole image by averaging on the overlapping areas. The coupled super-resolution algorithm is described in Algorithm 2.

IV. EXPERIMENTS

We now present a series of experiments to validate the effectiveness of the proposed joint image SR approach in various scenarios. In subsection IV-A, we perform multi-spectral image super-resolution (MS-SR) aided by the corresponding RGB version of the same scene. In subsection IV-B, we perform near-infrared image super-resolution (NIR-SR) aided by the corresponding RGB version of the same scene. We consider situations where the training and/or testing images are contaminated by noise in subsection IV-C to demonstrate the robustness of the proposed approach in comparison with other state-of-the-art approaches.

We compare our approach with state-of-the-art joint image filtering approaches, including Joint Bilateral Filtering (JBF) [22], Guided image Filtering (GF) [23], Static/Dynamic Filtering (SDF) [24], Deep Joint image Filtering (DJF) [25] and Joint Filtering via optimizing a Scale Map (JFSM) [26] where the same RGB guidance images as in our approach are leveraged. Our approach is also compared with several representative single image SR approaches, such as A+ [13], ANR (Anchored neighbourhood regression) [11], and the sparse coding algorithm of Zeyde *et al.* [10]. Furthermore,

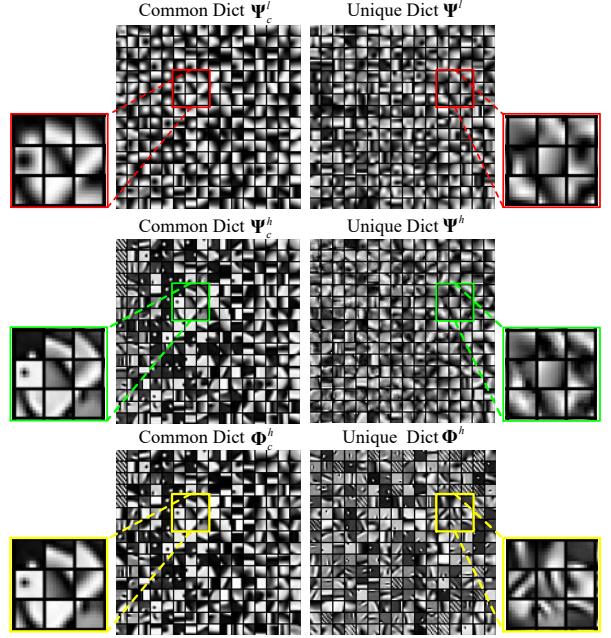


Figure 2: Learned coupled dictionaries for multi-spectral images of wavelength 640nm and RGB images. 256 atoms are shown here. The first row indicates the common and unique dictionaries learned for $4 \times$ downsampling LR multi-spectral images. The second row indicates the HR dictionary pair. The last row shows the dictionaries learned from side information, i.e. RGB images.

we select bicubic interpolation as the baseline method. We adopt the Peak Signal to Noise Ratio (PSNR), the root-mean-square error (RMSE) and the Structure SIMilarity (SSIM) index [48] as the image quality evaluation metrics which are commonly used in the image processing literature. The multi-spectral/RGB datasets are obtained from the Columbia multi-spectral database⁷. The infrared/RGB images datasets are obtained from the EPFL RGB-NIR Scene database⁸. All these datasets are registered for both modalities. For each multimodal dataset, we randomly separate its image pairs into two groups: training group and testing group. Then, we blur and downsample each HR image of target modality by a factor, e.g., $4 \times$ and $6 \times$, using the MATLAB "imresize" function to generate corresponding LR versions, similar to [8], [31].

A. Multi-spectral image SR

Training Phase with CDL. Before the coupled dictionary learning, we adopt some common preprocessing operations. Specifically, we upscale the LR multi-spectral training images to the desired size (i.e. the same size as HR version) using bicubic interpolation. The RGB images are converted from RGB to YCbCr space where we only use the luminance channel as the guidance, since human eyes are more sensitive to luminance information than chrominance information. Then, the interpolated LR images, the target HR images and the corresponding guidance images are divided into a set of

⁷<http://www.cs.columbia.edu/CAVE/databases/multispectral/>

⁸http://ivrl.epfl.ch/supplementary_material/cvpr11/

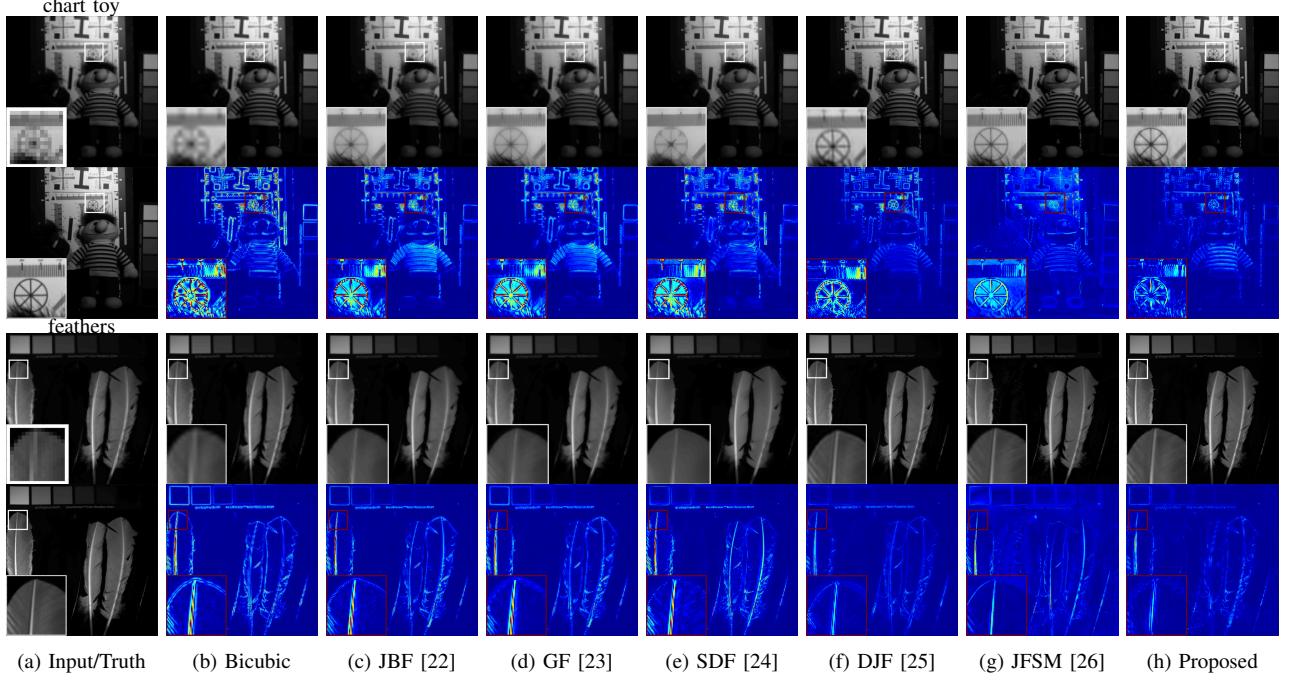


Figure 3: $4\times$ upscaling for multi-spectral images of 640nm wavelength. For each image, the first row is the LR input and SR results. The second row is the ground truth and corresponding error map for each approach. In the error map, brighter area represents larger error.

Table I: $4\times$ upscaling for multi-spectral image of 640 nm band evaluated by PSNR (dB) and SSIM

	Bicubic		JBF [22]		GF [23]		SDF [24]		DJF [25]		JFSM [26]		Proposed	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
chart toy	0.9451	29.14	0.9528	30.69	0.9514	30.70	0.9523	30.74	0.9842	33.91	0.9215	33.300	0.9855	34.50
cloth	0.7571	26.91	0.7640	27.62	0.7699	27.79	0.7315	27.18	0.9489	31.54	0.9770	35.330	0.9506	32.75
egyptian	0.9761	36.22	0.9788	37.82	0.9788	37.96	0.9677	37.16	0.9861	41.31	0.9428	39.680	0.9935	42.63
feathers	0.9530	30.46	0.9599	31.80	0.9618	32.12	0.9434	30.92	0.9848	36.01	0.9096	33.540	0.9871	36.25
glass tiles	0.9215	26.38	0.9339	27.15	0.9326	27.45	0.9188	27.01	0.9814	31.83	0.9407	29.340	0.9791	31.05
jelly beans	0.9269	27.45	0.9474	28.97	0.9488	29.54	0.9279	27.87	0.9820	32.77	0.9356	30.820	0.9866	34.38
oil painting	0.9025	32.23	0.9034	33.23	0.9033	33.30	0.9001	32.80	0.9493	34.39	0.9439	34.160	0.9601	36.24
paints	0.9569	30.47	0.9714	32.08	0.9698	32.23	0.9569	31.35	0.9897	37.74	0.9321	32.960	0.9900	36.99
average	0.9174	29.91	0.9265	31.17	0.9270	31.39	0.9123	30.63	0.9758	34.94	0.9379	33.640	0.9791	35.60

$\sqrt{N} \times \sqrt{N}$ patch pairs. We remove the mean from each patch, as the DC component is always preserved well during the upscaling process. Then, we vectorize the patches to form the training datasets \mathbf{X}^l , \mathbf{X}^h and \mathbf{Y} of dimension $N \times T$. Smooth patches with variance less than 0.02 have been eliminated as they are less informative. Once the training dataset is prepared, we apply our coupled dictionary learning algorithm, shown in Algorithm 1, to learn the dictionary pairs $[\Psi_c^l, \Psi_c^l]$ and $[\Phi_c, \Phi]$ from \mathbf{X}^l and \mathbf{Y} . Then, HR dictionary pair $[\Psi_c^h, \Psi_c^h]$ are computed based on \mathbf{X}^h and the acquired sparse codes \mathbf{Z} and \mathbf{U} . The parameter setting is as follows: patch size $\sqrt{N} \times \sqrt{N} = 8 \times 8$ for $4\times$ upscaling and 16×16 for $6\times$ upscaling, dictionary size $K = 1024$, total sparsity constraint $s = 20$, training size $T \approx 15,000$.

Figure 2 shows the learned coupled dictionaries for multi-spectral images of wavelength 640 nm and the corresponding RGB version. We can find that any pair of LR and HR atoms from Ψ_c^l and Ψ_c^h capture associated edges, blobs, textures with the same direction and location. Similar behavior can also be observed in Ψ^l and Ψ^h . This implies that LR and HR dictionaries are indeed closely related to each other. On the other

hand, LR and HR atom pairs also exhibit some differences. Specifically, the edges and textures captured by LR atoms tend to be blurred and smoothed, while they tend to be clearer and sharper in the corresponding HR atoms. More importantly, the common dictionary Φ_c^h from the guidance images exhibits considerable resemblance and strong correlation to Ψ_c^h and Ψ_c^l from the HR/LR modalities of interest. This indicates that the three common dictionaries have indeed captured the similarities between multi-spectral and RGB modalities. In contrast, the learned unique dictionaries Ψ^h and Φ represent the disparities of these modalities and therefore rarely exhibit resemblance.

Testing Phase with CSR. During the coupled super-resolution phase, given a new pair of LR multi-spectral and HR RGB images for test, we upscale the LR multi-spectral image to the desired size as before. Then the testing image pair are subdivided into overlapping patches of size $\sqrt{N} \times \sqrt{N}$ pixels with overlap stride equal to 1 pixel.⁹ The DC component is also removed from each patch and stored. We vectorize

⁹The overlap stride denotes the distance between corresponding pixel locations in adjacent image patches.

Table II: $6\times$ upscaling for multi-spectral image of 640 nm band evaluated by PSNR (dB) and SSIM

	Bicubic		JBF [22]		GF [23]		SDF [24]		DJF [25]		JFSM [26]		Proposed	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
chart toy	0.8774	26.83	0.8992	28.08	0.8932	27.86	0.9006	28.12	0.9772	32.61	0.9144	31.22	0.9682	32.55
cloth	0.6143	25.55	0.6424	26.06	0.6394	26.07	0.6158	25.80	0.9226	30.09	0.9723	33.79	0.9256	31.73
egyptian	0.9459	33.79	0.9560	34.95	0.9536	34.80	0.9466	34.83	0.9681	40.24	0.9444	38.43	0.9872	40.75
feathers	0.8973	27.68	0.9177	28.80	0.9138	28.76	0.9062	28.50	0.9765	34.09	0.9042	31.32	0.9727	33.75
glass tiles	0.8401	24.45	0.8652	25.05	0.8585	25.06	0.8556	25.03	0.9705	30.33	0.9233	27.33	0.9646	29.87
jelly beans	0.8424	24.93	0.8835	26.24	0.8801	26.36	0.8681	25.60	0.9721	31.22	0.9225	28.58	0.9734	32.73
oil painting	0.8511	30.90	0.8664	31.87	0.8626	31.78	0.8574	31.49	0.9392	33.77	0.9462	34.09	0.9427	35.18
paints	0.9005	27.51	0.9328	29.04	0.9253	28.90	0.9226	28.69	0.9842	36.60	0.9363	31.25	0.9792	34.93
average	0.8461	27.70	0.8704	28.76	0.8658	28.70	0.8591	28.50	0.9638	33.62	0.9330	32.00	0.9642	33.93

these patches to construct the testing datasets \mathbf{x}_{test}^l and \mathbf{y}_{test} . Then, we perform coupled sparse coding on \mathbf{x}_{test}^l and \mathbf{y}_{test} with respect to learned dictionary pairs $[\Psi_c^l, \Psi^l]$ and $[\Phi_c, \Phi]$ to obtain the approximated sparse codes \mathbf{z}_{test} , \mathbf{u}_{test} and \mathbf{v}_{test} , which are then multiplied with the HR dictionary pair $[\Psi_c^h, \Psi^h]$ to predict the HR patches \mathbf{x}_{test}^h , shown in Algorithm 2. Finally, the DC component of each patch is added back to the corresponding estimated HR patch. These HR patches are tiled together and the overlapping areas are averaged to reconstruct the HR image of interest.

Figure 3 shows the multi-spectral image SR results for the 640 nm wavelength band. As we can see, the reconstructed MS image and its corresponding residual from bicubic interpolation, JBF [22], GF [23], SDF [24] and DJF [25] exhibit noticeable blurred areas. The reconstruction from JFSM [26] shows sharp edges but with weaker intensity than the ground-truth, a form of luminance distortion resulting from texture copying artifacts (see the zoom-in area of the wheel in the chart toy). In comparison, our approach is able to reliably recover more accurate image details and, at the same time, substantially suppresses ringing artifacts. Therefore, our reconstruction is more photo-realistic and visually appealing than the counterparts. This is also confirmed by the error maps, as well as by quantitative measure in terms of PSNR and SSIM, shown in Table I and Table II for $4\times$ and $6\times$ upscaling, respectively.¹⁰ The quantitative results show that our approach outperforms bicubic interpolation with significant gains of average 5.6dB, 6.2dB and also exhibits notable advantage over the state-of-the-art joint image filtering approaches. For both $4\times$ and $6\times$ upscaling, the proposed approach outperforms JBF [22], GF [23], SDF [24], JFSM [26] with gains of at least 1.9dB in terms of average PSNR. Our approach also outperforms the deep-learning-based approach DJF [25] for the selected number of training samples.

B. Near-infrared image SR

We also evaluate our approach on near-infrared (NIR) images with registered RGB images as side information. As the response of NIR band has poor correlation with the response of the visible band, it is usually difficult to infer the brightness of a NIR image given a corresponding RGB modality. Thus, it is more challenging to take good advantage of the RGB version to super-resolve the near-infrared version. The LR/HR training/testing dataset and the side information

¹⁰Limited to space, only a few algorithms producing the best results are shown in the paper. More detailed results can be found in the supplementary materials.

are prepared in a manner similar to the previous multi-spectral case. The parameter setting keeps the same as before. The first dataset includes houses and buildings that contain many fine textures and sharp edges. This makes the SR task more challenging than super-resolving images with smoother textures. The second dataset includes natural landscape images with water, trees, stone and more.

Figure 4 compares the visual quality of the reconstructed HR near-infrared images and the corresponding error maps. It can be seen that, on average, our approach recovers more visually plausible images, exhibiting less error than the competing methods. Table III and IV also confirm the significant advantage of the proposed approach over other state-of-the-art methods. In particular, this indicates that detailed structure information can be effectively captured by coupled dictionary learning, especially on images such as buildings and houses that contain a lot of sharp edges, textures and stripes.

Figure 5 and Table V show the visual and quantitative comparison for another dataset with landscape images. It can be seen that leaves, trees, grass and other natural objects with fine details tend to be over-smoothed in the reconstructed images from competing approaches. In contrast, these objects in our reconstruction appear clearer, sharper and less obscured. This further confirms the advantage of CDLSR in reliably restoring fine details without introducing notable artifacts. (See additional comparisons with DJF [25] in subsection IV-C.)

Overall, the good performance of the proposed CDLSR approach is due to learned adaptive coupled dictionaries that are capable of effectively capturing salient features and critical correlations between the target and the guidance modalities in their sparse transform domains. These learned dictionaries can act as powerful priors that have the ability to dramatically reduce artifacts.

C. Proposed CDLSR vs Deep-Learning-Based SR Approaches

The previous experiments have shown that for a relatively modest number of training samples our proposed approach can lead to better results than the state-of-the-art, including deep-learning-based multimodal super-resolution methods (DJF [25]). However, as deep-learning-based methods can also successfully take advantage of the availability of a huge amount of data for training, DJF [25] eventually outperforms our approach, given enough training data.

Our proposed approach has nonetheless other advantages with respect to DJF [25]. One advantage relates to the amount of training time required by DJF [25] in relation to CDLSR. For example, DJF [25] takes about 12 hours to train through

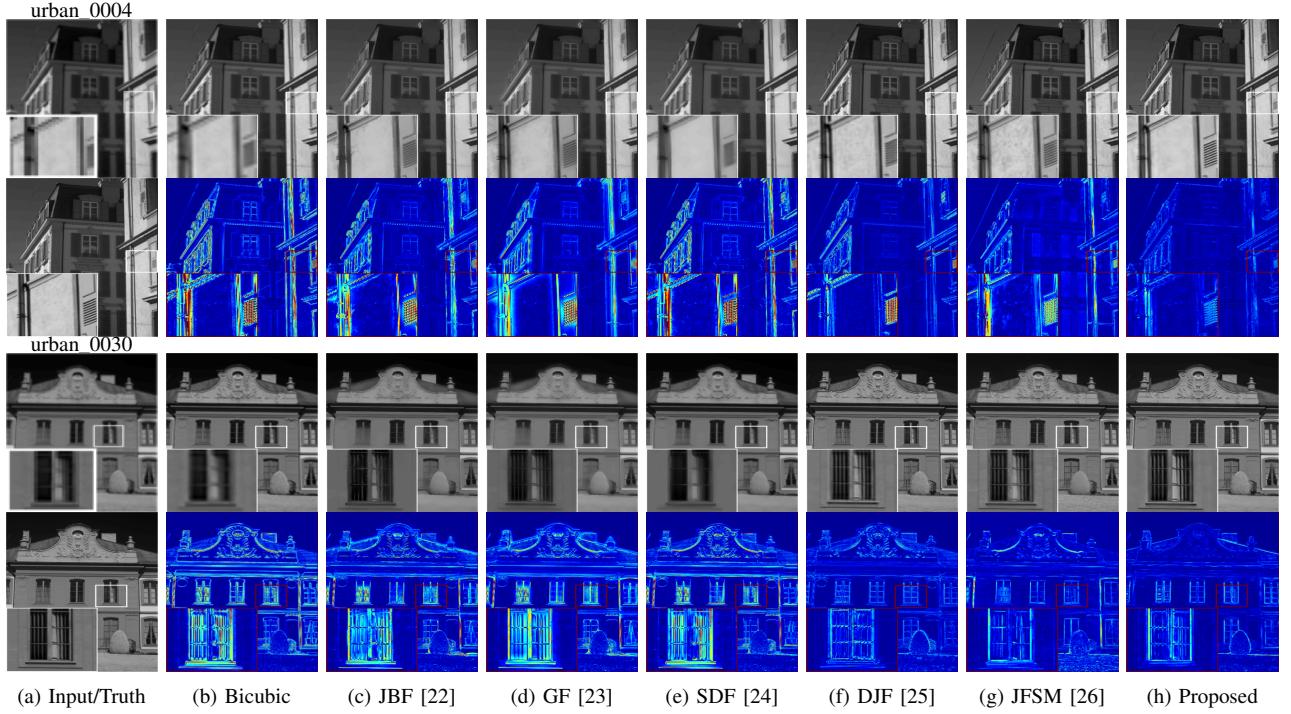


Figure 4: $4\times$ upscaling for near-infrared house images, e.g., urban_0004(up) and urban_0030(bottom). For each image, the first row is the LR input and SR results. The second row is the ground truth and corresponding error map for each approach. In the error map, brighter area represents larger error.

Table III: $4\times$ upscaling for near-infrared house images

	Bicubic		JBF [22]		GF [23]		SDF [24]		DJF [25]		JFSM [26]		Proposed	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.9029	25.93	0.9359	28.47	0.9391	28.75	0.9066	26.82	0.9789	31.02	0.9721	30.86	0.9811	34.14
urban_0006	0.9458	30.89	0.9311	32.10	0.9400	32.66	0.8918	30.60	0.9894	36.04	0.9741	32.86	0.9868	36.79
urban_0017	0.9527	30.45	0.9172	31.11	0.9205	31.32	0.9281	30.72	0.9815	34.18	0.9500	32.85	0.9777	35.27
urban_0018	0.9298	25.19	0.9308	27.59	0.9251	27.70	0.9196	26.09	0.9888	30.72	0.9774	30.80	0.9874	33.01
urban_0020	0.9577	28.03	0.9523	30.67	0.9494	30.69	0.9505	29.09	0.9915	33.60	0.9797	32.61	0.9893	36.66
urban_0026	0.8704	26.27	0.8627	26.82	0.8571	26.89	0.8558	26.61	0.9397	29.21	0.9332	28.97	0.9482	30.35
urban_0030	0.8401	26.54	0.8476	27.58	0.8383	27.59	0.8415	27.21	0.9345	31.27	0.9064	30.56	0.9443	32.71
urban_0050	0.9434	26.65	0.9099	27.32	0.9116	27.35	0.9207	27.07	0.9616	28.58	0.9251	27.58	0.9663	29.37
average	0.9179	27.49	0.9109	28.96	0.9101	29.12	0.9018	28.03	0.9707	31.83	0.9522	30.89	0.9726	33.54

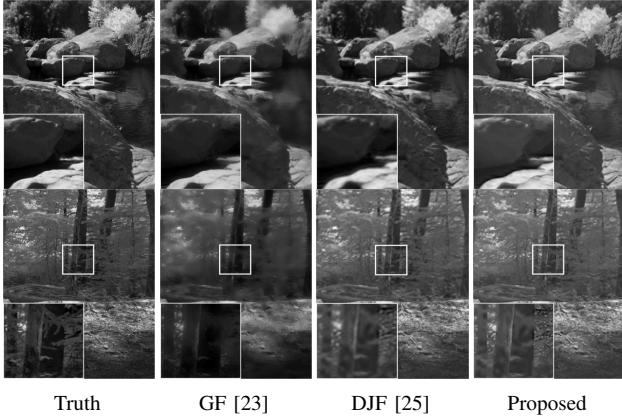


Figure 5: $4\times$ upscaling for near-infrared landscape images, e.g., n0031(up) and n0051(bottom).

50 epochs with an NVIDIA Titan black GPU for acceleration, while our approach takes only a few minutes for training a group of coupled dictionaries without any GPU acceleration. But our approach is slower than the deep-learning-based

approach DJF during testing because we solve a non-convex optimization problem while DJF only performs a simple forward pass.

More importantly, the other advantage relates to the robustness of our approach in the presence of noise at training and/or testing stages, which is very common in practice [49], [50]. In particular, we repeat the previous NIR-SR experiments to test the robustness of both algorithms in the presence of contamination of additive zero-mean Gaussian noise. The training dataset for DJF [25] consists of 160,000 33×33 patches and for the proposed CDLSR only 15000 patches of size 8×8 pixels. Each evaluation metric value is averaged on all the testing images. We consider two typical scenarios:¹¹

1) *LR noisy testing images*: The first scenario assumes that the LR testing images are contaminated by zero-mean Gaussian noise with a certain standard deviation. Note that coupled dictionary learning is conducted on noiseless training

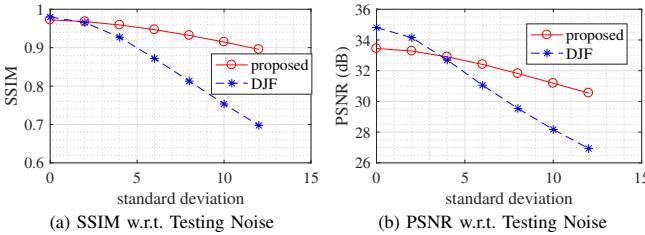
¹¹We assume that only the target modality is contaminated by noise and the guidance modality keeps clean as before in order to compare with previous noise-free situations.

Table IV: $6\times$ upscaling for near-infrared house images

	Bicubic		JBF [22]		GF [23]		SDF [24]		DJF [25]		JFSM [26]		Proposed	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.8094	23.87	0.8858	25.96	0.8817	25.94	0.8413	24.62	0.9670	29.68	0.9527	27.97	0.9558	30.77
urban_0006	0.8671	28.48	0.8861	30.00	0.8876	30.17	0.8377	28.76	0.9830	34.92	0.9716	32.28	0.9664	34.15
urban_0017	0.8998	28.64	0.8864	29.63	0.8860	29.61	0.8910	29.13	0.9599	32.80	0.9434	32.01	0.9515	32.98
urban_0018	0.8393	23.07	0.8718	25.09	0.8591	24.98	0.8439	23.79	0.9844	29.92	0.9470	27.47	0.9727	31.03
urban_0020	0.9053	26.03	0.9200	28.19	0.9118	28.01	0.9089	26.93	0.9873	32.61	0.9673	30.33	0.9763	33.85
urban_0026	0.7850	24.71	0.8235	25.64	0.8131	25.63	0.7989	25.17	0.9183	28.38	0.9128	27.54	0.9172	28.88
urban_0030	0.7517	25.19	0.7994	26.32	0.7855	26.22	0.7748	25.80	0.9063	30.00	0.8902	29.38	0.9099	30.52
urban_0050	0.8921	25.17	0.8837	26.26	0.8846	26.26	0.8837	25.90	0.9414	27.64	0.9068	26.67	0.9402	28.37
average	0.8437	25.65	0.8696	27.13	0.8637	27.10	0.8475	26.26	0.9559	30.75	0.9365	29.21	0.9487	31.32

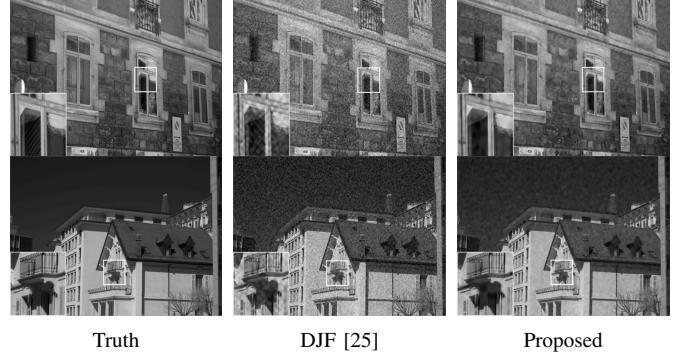
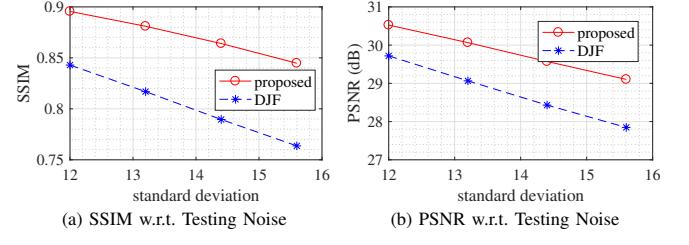
Table V: $4\times$ upscaling for near-infrared landscape images.

	GF [23]		JFSM [26]		DJF [25]		Proposed	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
n0025	0.7970	27.14	0.8242	25.67	0.9093	28.43	0.9097	29.05
n0027	0.7002	25.82	0.7033	24.68	0.8565	27.87	0.8702	28.07
n0028	0.7519	25.01	0.7766	24.16	0.8812	26.88	0.8789	26.50
n0031	0.8524	27.81	0.8536	26.71	0.9111	28.72	0.9136	28.64
n0049	0.7832	29.52	0.7453	26.85	0.9021	31.49	0.8996	31.88
n0051	0.7262	25.97	0.7606	25.19	0.8732	27.64	0.8767	28.29
average	0.7685	26.88	0.7773	25.54	0.8889	28.50	0.8914	28.74

Figure 6: LR noisy testing images. The testing noise σ_{test} ranges from 2 to 12 and the training noise $\sigma_{train} = 0$, for $4\times$ upscaling of near-infrared house images using DJF [25] and proposed CDLSR.

images but coupled image super-resolution is conducted on the noisy LR images. In Table VI and Figure 6, the results corresponding to setting $\sigma_{test} = \sigma_{train} = 0$ show that deep-learning-based multimodal super-resolution method DJF [25] usually outperforms our approach given a large number of training samples in the noise-free scenario. However, other results corresponding to setting $\sigma_{test} \neq 0$ show that our proposed algorithm demonstrates reasonable stability and robustness to noise, especially to strong noise. In contrast, DJF [25] is susceptible to noise and its performance degrades faster than ours. In Figure 7, it can be observed that the upscale results of DJF [25] can not attenuate noise effectively, whereas our reconstruction is much cleaner. We believe that the good robustness and stability is due to sparsity priors exploited by our model.

2) *LR noisy both testing and training images*: The second scenario assumes that both the testing and the training images are contaminated by zero-mean Gaussian noise with a certain standard deviation. Coupled dictionary learning is performed on the noisy training images and coupled image super-resolution is done on the noisy testing image. In addition, we consider possible *mismatch* of noise in the LR testing and training images as well. Specifically, given a certain standard deviation σ_{train} for the training noise and mismatch δ , the standard deviation of the corresponding testing noise is set as $\sigma_{test} = \sigma_{train}(1 + \delta)$. We add noise with standard deviation

Figure 7: LR noisy testing images. The testing noise $\sigma_{test} = 12$ and the training noise $\sigma_{train} = 0$, for $4\times$ upscaling of urban_0006(up) and urban_0020(bottom) using DJF [25] and proposed CDLSR.Figure 8: LR noisy both testing and training images. $\sigma_{train} = 12$ and σ_{test} ranges from 12 to 15.6. $4\times$ upscaling of near-infrared house images using DJF [25] and proposed CDLSR.

σ_{train} in the LR training images and noise with standard deviation σ_{test} in the LR testing images for various values of δ . For example, given a typical noise level $\sigma_{train} = 12$ and mismatch $\delta = [0, 10\%, 20\%, 30\%]$, it leads to corresponding $\sigma_{test} = [12, 13.2, 14.4, 15.6]$. Then we repeat the previous training and testing for $4\times$ upscaling of near-infrared house images using both CDLSR and DJF [25]. As shown in Table VII and Figure 8, the performance of both the proposed CDLSR approach and DJF [25] degrades as the mismatch increases. However, the proposed algorithm not only has a slower degradation in performance than DJF [25], but also yields higher SSIM and PSNR values. This illustrates that our method is more robust to mismatched noise.

D. Impact of parameters

In this section, we illuminate further the performance of the CDLSR algorithm by exploring the effect of key parameters and factors on the performance of the proposed algorithms. The following experiments are conducted using multi-spectral

Table VI: LR noisy testing images. The standard deviation of the testing noise is $\sigma_{test} = [0, 4, 8, 12]$ while the standard deviation of the training noise is $\sigma_{train} = 0$, for $4 \times$ upscaling of near-infrared house images using DJF [25] and CDLSR.

	$\sigma_{train} = 0, \sigma_{test} = 0$				$\sigma_{train} = 0, \sigma_{test} = 4$				$\sigma_{train} = 0, \sigma_{test} = 8$				$\sigma_{train} = 0, \sigma_{test} = 12$			
	Proposed		DJF [25]		Proposed		DJF [25]		Proposed		DJF [25]		Proposed		DJF [25]	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.9811	34.14	0.9895	35.85	0.9715	32.96	0.9386	33.62	0.9477	31.97	0.8313	30.21	0.9150	30.71	0.7181	27.42
urban_0006	0.9868	36.79	0.9917	37.92	0.9752	35.77	0.9524	34.54	0.9483	33.71	0.8574	30.33	0.9132	31.87	0.7493	27.23
urban_0017	0.9777	35.27	0.9861	37.00	0.9592	34.62	0.9233	34.19	0.9243	33.08	0.7896	30.30	0.8777	31.47	0.6571	27.33
urban_0018	0.9874	33.01	0.9933	34.48	0.9801	32.58	0.9537	32.75	0.9594	31.88	0.8650	29.75	0.9302	30.72	0.7766	27.13
urban_0020	0.9893	36.66	0.9953	38.27	0.9784	36.11	0.9334	34.94	0.9479	34.51	0.8070	30.74	0.9063	32.64	0.6807	27.63
urban_0026	0.9482	30.35	0.9635	31.53	0.9339	30.00	0.9105	30.58	0.9104	29.46	0.7968	28.54	0.8788	28.69	0.6808	26.44
urban_0030	0.9443	32.71	0.9604	35.10	0.9278	32.25	0.9101	33.09	0.9043	31.37	0.8003	29.81	0.8754	30.29	0.6907	27.06
urban_0050	0.9663	29.37	0.9586	28.31	0.9465	28.92	0.8935	27.84	0.9138	28.54	0.7559	26.66	0.8687	28.01	0.6270	25.23
average	0.9726	33.54	0.9798	34.81	0.9591	32.90	0.9269	32.69	0.9320	31.82	0.8129	29.54	0.8956	30.55	0.6975	26.93

Table VII: LR noisy both testing and training images. The standard deviation of the training noise is $\sigma_{train} = 12$ and the standard deviation of the testing noise σ_{test} ranges from 12 to 15.6, corresponding to mismatch δ ranging from 0 to 30%. $4 \times$ upscaling of near-infrared house images using DJF [25] and proposed CDLSR.

	$\sigma_{train} = 12, \sigma_{test} = 12$				$\sigma_{train} = 12, \sigma_{test} = 13.2$				$\sigma_{train} = 12, \sigma_{test} = 14.4$				$\sigma_{train} = 12, \sigma_{test} = 15.6$			
	Proposed		DJF [25]		Proposed		DJF [25]		Proposed		DJF [25]		Proposed		DJF [25]	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.9148	30.67	0.8724	30.38	0.9029	30.21	0.8545	29.84	0.8884	29.73	0.8233	29.03	0.8716	29.28	0.7981	28.38
urban_0006	0.9131	31.82	0.8746	30.57	0.8973	31.17	0.8497	29.77	0.8789	30.48	0.8264	29.05	0.8585	29.84	0.8022	28.33
urban_0017	0.8777	31.46	0.8141	30.63	0.8606	30.98	0.7792	29.71	0.8412	30.46	0.7474	29.07	0.8188	29.93	0.7156	28.40
urban_0018	0.9301	30.70	0.8903	29.97	0.9174	30.20	0.8713	29.39	0.9029	29.67	0.8479	28.70	0.8870	29.19	0.8306	28.16
urban_0020	0.9062	32.58	0.8442	31.30	0.8895	31.94	0.8161	30.50	0.8704	31.27	0.7824	29.66	0.8492	30.68	0.7608	29.04
urban_0026	0.8786	28.66	0.8344	28.60	0.8661	28.36	0.8097	28.08	0.8511	28.02	0.7877	27.67	0.8341	27.73	0.7606	27.19
urban_0030	0.8754	30.29	0.8275	29.83	0.8629	29.85	0.8007	29.09	0.8483	29.39	0.7785	28.44	0.8313	28.90	0.7498	27.76
urban_0050	0.8687	28.01	0.7860	26.47	0.8514	27.79	0.7539	26.14	0.8311	27.55	0.7229	25.83	0.8076	27.27	0.6915	25.53
average	0.8956	30.52	0.8429	29.72	0.8810	30.07	0.8169	29.06	0.8640	29.57	0.7896	28.43	0.8448	29.10	0.7636	27.85

Table VIII: Effect of dictionary size

# of atoms	64	128	256	512	1024	2048
Train Time	1.8	5.5	10.0	21.9	62.7	253.8
Test Time	92.5	95.6	117.6	138.8	142.0	216.3
PSNR (dB)	33.81	34.38	34.57	34.83	34.95	35.33
RMSE	0.0204	0.0191	0.0187	0.0181	0.0179	0.0171

images of wavelength 640 nm on which we perform $4 \times$ upscaling using a computer equipped with a quadro-core i7 CPU at 3.4GHz with 32GB of memory. Each evaluation metric value is averaged on all the testing images.

Dictionary Size. Intuitively, more atoms tend to capture more features. Thus, a larger dictionary may yield a more accurate sparse approximation to the signal of interest. On the other hand, a large dictionary size increases the complexity of the non-convex problem, thus requiring more computation. Under the multi-spectral image SR experimental setting, we evaluate the performance of the proposed approach for various dictionary sizes, including 64, to 128, 256, 512, and 1024 atoms. Table VIII shows that the PSNR increases gradually with the increase of the dictionary size. On the other hand, the computation cost, represented by the training time and testing time, approximately increases linearly with the dictionary size. The results imply that the choice of the dictionary size depends on the balance between approximation accuracy and computational expense.

Sparsity Constraints. A larger sparsity constraint, i.e., a larger s , can lead to a better approximation of the data. On the other hand, larger sparsity constraints also require more iterations to find these non-zeros via OMP. As shown in Table IX, the average PSNR of the reconstruction, as well as the computational time, increase along with the total sparsity

Table IX: Effect of sparsity constraints

total sparsity	8	12	16	20	24	28
Test Time (s)	5.7	8.8	12.6	17.9	23.2	31.9
PSNR (dB)	34.97	35.54	35.69	35.73	35.73	35.69

constraint. When the total sparsity constraint goes beyond a certain level, e.g., 16, the retrieved extra non-zeros coefficients are trivial and contribute very little to the PSNR.

These considerations suggest that a dictionary size around 1024 atoms with sparsity constraint around 20 for 8×8 image patch size can yield decent performance while allowing affordable computational complexity.

V. CONCLUSION

This paper proposed a new multimodal image SR approach based on joint sparse representations and coupled dictionary learning. In particular, our CDLSR approach explicitly captures the similarities and disparities between different image modalities in the sparse feature domain in *lieu* of the image domain. The proposed CDLSR approach consists of a training phase and a testing phase. The training phase seeks to learn a number of coupled dictionaries from training data and the testing phase leverages the learned dictionaries to reconstruct a HR version of a LR image with the aid of the guidance image. Our design automatically transfers appropriate structure information to the estimated HR version. Multispectral/RGB and NIR/RGB multimodal image SR experiments demonstrate that our design brings notable benefits over state-of-the-art image SR approaches. Our approach also outperforms deep-learning-based methods especially when the data is contaminated by noise, demonstrating better robustness, but consuming much less computing resource and training time.

REFERENCES

- [1] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, 2001.
- [2] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* IEEE, 2007, pp. 1–8.
- [3] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* IEEE, 2008, pp. 1–8.
- [4] X. Zhang and X. Wu, "Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 887–896, 2008.
- [5] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* IEEE, 2013, pp. 1059–1066.
- [6] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, 2013.
- [7] S. Mallat and G. Yu, "Super-resolution with sparse mixing estimators," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2889–2900, 2010.
- [8] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [9] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, 2012.
- [10] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*. Springer, 2010, pp. 711–730.
- [11] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vision*. IEEE, 2013, pp. 1920–1927.
- [12] X. Wei and P. L. Dragotti, "Fresh – fri-based single-image super-resolution algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3723–3735, 2016.
- [13] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vision*. Springer, 2014, pp. 111–126.
- [14] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.
- [15] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2016, pp. 391–407.
- [16] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1646–1654.
- [17] J. Kim, J. Kwon Lee *et al.*, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1637–1645.
- [18] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: a review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.
- [19] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanusot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes *et al.*, "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, 2015.
- [20] F. Renna, L. Wang, X. Yuan, J. Yang, G. Reeves, R. Calderbank, L. Carin, and M. R. Rodrigues, "Classification and reconstruction of high-dimensional signals from low-dimensional features in the presence of side information," *IEEE Trans. Inform. Theory*, vol. 62, no. 11, pp. 6459–6492, 2016.
- [21] J. F. Mota, N. Deligiannis, and M. R. Rodrigues, "Compressed sensing with prior information: Strategies, geometry, and bounds," *IEEE Trans. Inform. Theory*, vol. 63, no. 7, 2017.
- [22] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *ACM Trans. Graph.*, vol. 26, no. 3. ACM, 2007, p. 96.
- [23] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2010, pp. 1–14.
- [24] B. Ham, M. Cho, and J. Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
- [25] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2016, pp. 154–169.
- [26] X. Shen, Q. Yan, L. Xu, L. Ma, and J. Jia, "Multispectral joint image restoration via optimizing a scale map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2518–2530, 2015.
- [27] Q. Zhang, X. Shen, L. Xu, and J. Jia, "Rolling guidance filter," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2014, pp. 815–830.
- [28] P. Song, J. F. Mota, N. Deligiannis, and M. R. Rodrigues, "Coupled dictionary learning for multimodal image super-resolution," in *IEEE Global Conf. Signal Inform. Process.* IEEE, 2016, pp. 162–166.
- [29] S. Hawe, M. Kleinsteuber, and K. Diepold, "Analysis operator learning and its application to image reconstruction," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2138–2150, 2013.
- [30] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.
- [31] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* IEEE, 2008, pp. 1–8.
- [32] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* IEEE, 2012, pp. 2216–2223.
- [33] K. Jia, X. Wang, and X. Tang, "Image transformation based on learning dictionaries across image spaces," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 367–380, 2013.
- [34] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [35] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [36] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Int. Conf. Comput. Vision*. IEEE, 1998, pp. 839–846.
- [37] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.
- [38] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2014, pp. 3550–3557.
- [39] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2015, pp. 695–704.
- [40] M. Dao, N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Collaborative multi-sensor classification via sparsity-based representation," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2400–2415, 2016.
- [41] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 24–38, 2016.
- [42] N. Deligiannis, J. F. Mota, B. Cornelis, M. R. Rodrigues, and I. Daubechies, "Multi-modal dictionary learning for image separation with application in art investigation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 751–764, 2017.
- [43] N. Deligiannis, J. F. Mota, B. Cornelis, M. R. Rodrigues *et al.*, "X-ray image separation via coupled dictionary learning," in *IEEE Int. Conf. Image Process.* IEEE, 2016, pp. 3533–3537.
- [44] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [45] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [46] J. Wang, S. Zhu, and Y. Gong, "Resolution enhancement based on learning the sparse association of image patches," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 1–10, 2010.
- [47] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.
- [48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [49] L. Zhang, W. Dong, D. Zhang, and G. Shi, "Two-stage image denoising by principal component analysis with local pixel grouping," *Pattern Recognition*, vol. 43, no. 4, pp. 1531–1549, 2010.
- [50] M. P. Nguyen and S. Y. Chun, "Bounded self-weights estimation method for non-local means image denoising using minimax estimators," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1637–1649, 2017.