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 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 guidance, based on joint sparse representations induced by coupled dictionaries. The proposed approach captures complex dependency correlations, including similarities and disparities, between different image modalities in a learned sparse feature domain in lieu of the original image domain. It consists of two phases: coupled dictionary learning phase and coupled super-resolution phase. The learning phase learns a set of dictionaries from the training dataset to couple different image modalities together in the sparse feature domain. In turn, the super-resolution phase leverages such dictionaries to construct a HR version of the LR target image with another related image modality for guidance. In the advanced version of our approach, multi-stage strategy and neighbourhood regression concept are introduced to further improve the model capacity and performance. Extensive guided image super-resolution experiments on real multimodal images demonstrate that the proposed approach admits distinctive advantages with respect to the state-of-the-art approaches, for example, overcoming the texture copying artifacts commonly resulting from inconsistency between the guidance and target images. Of particular relevance, the proposed model demonstrates much better robustness than competing deep models in a range of noisy scenarios.**

***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 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]–[18]. These typical super-resolution approaches focus only on single modality images without exploiting the availability of additional modalities as guidance.

However, in many practical application scenarios, a certain scene is often imaged using different sensors to yield 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 [19], [20]. In order to balance cost, bandwidth and complexity, these multimodal images are usually acquired with different resolutions [19]. 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* [21], [22], to aid the super-resolution of target LR modalities [23]–[29]. 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 Coupled Dictionary Learning based multimodal image Super-Resolution approach, termed as CDLSR. Our approach exploits joint sparse representations induced by coupled dictionaries to capture complex dependencies between different modalities in a learned sparse feature domain. It also takes into account both similarities and disparities between target and guidance images to avoid introducing noticeable texture-copying artifacts.

Proposed Scheme. The proposed scheme is based on three elements: a data model, a coupled dictionary learning

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: ndeligia@etrovub.be)

These authors contributed equally: Pingfan Song, Xin Deng.

Code and data is available on https://github.com/pingfansong/CDLSR

algorithm and 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 image patches of different modalities. Of particular relevance is the ability to capture the complex dependency correlations among these modalities in a learned sparse feature domain in lieu of the original image domain.
- *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 representative guided/joint filtering methods [23]–[26] that exploit "shallow" models for guided super-resolution, our learning based approach can better model the complex dependencies in learned sparse domains. In addition, we also take both the common and distinct features of the different data modalities into consideration, instead of unilaterally transferring the structure details from the HR guidance image. This makes our approach more robust to inconsistencies between the guidance and the target images. Our approaches also demonstrate better robustness to (mismatched) noise and faster training speed in comparison to "deep models" [27], [28] in a range of scenarios.

Contributions. The contributions of this paper include:

- We devise a data model for multimodal signals to capture complex dependency correlations using joint sparse representations induced by coupled dictionaries. Compared with our previous work [30], the present model is more general, because it does not require the matrix that models the conversion of a HR image to its LR counterpart to be known.
- A learning algorithm is proposed to learn the coupled dictionaries from different data modalities. Again, compared with [30], in the learning stage, the proposed algorithm does not require the knowledge of the matrix that converts a HR image to a LR version.
- A multimodal image super-resolution algorithm is developed to enhance the resolution of the target LR image with the aid of another guidance HR image modality.
- Further, we present an advanced version – multi-stage CDLSR – where the proposed basic model is transformed into a deeper one consisting of multiple stages of coupled dictionary learning and coupled super-resolution operations. We also integrated neighbourhood regression concept to take better advantage of large amounts of training samples, further increasing the performance and speed of the proposed algorithm.
- Finally, extensive experiments are conducted on a variety of multimodal images. The results demonstrate that the proposed approach leads to better super-resolution performance and model robustness than the state-of-the-art.

Organization. The remainder of this paper is organized as follows. We review related work in Section II, including single and guided image SR, as well as other work in multimodal image processing. 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 presents an advanced version of the proposed model and algorithm. Section V is devoted to various practical experiments. We summarize and conclude in Section VI.

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 additional 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 [31], 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], [32]–[35]. For example, motivated by Compressive Sensing [36], [37], Yang *et al.* [8], [9], [33] 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 [34] and coupled dictionary learning [35] 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], [34], [35], 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. Liu *et al.* [18] proposes a Cascaded Sparse Coding Network (CSC-Net) to demonstrate that a sparse coding model particularly designed for SR can be incarnated as a neural network with the merit of end-to-end optimization over training data.

B. Guided image SR

Compared with single image SR, guided 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 [38] 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 pixel values themselves. It smoothes the image while preserving edges. The joint bilateral upsampling [23] 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 [24]. Later, guided image filtering [24] was proposed to overcome this limitation and is capable of preserving both edges and gradients using a simple dependency assumption (e.g. linear relationship between target and guidance images). To address notable appearance change problems that result from directly transferring gradients of guidance images, [26] proposes a framework that optimizes a novel scale map to capture the nature of structure discrepancy between images. If filters are designed based on the guidance image unilaterally, they may suffer from the inconsistency of the local structures in the guidance and target images, and thus transfer incorrect structure details to the target images. To this end, the study in [25] 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.

The aforementioned guided/joint filtering models are "shallow" ones with limited capacity, and they mainly employ hand-crafted features that may not reflect natural image priors well. Recent work [27] proposes a Convolutional Neural Network (CNN) based joint image filtering approach which leverages the guidance image as a prior and transfers the structural details from the guidance image to the target image for enhancing spatial resolution. Their design consists of three sub-networks of which two are used to extract informative features from the target and guidance image modalities and these features are then concatenated together by the third sub-network for final reconstruction. Gu *et al.* propose a weighted analysis sparse representation model (WASR) [28] which consists of multiple layers/stages of analysis filters to perform convolutional sparse coding in a way similar to the convolutional neural network. Their model deals with guided super-resolution tasks by integrating the guidance modality to generate specific weights. More detailed discussion to these methods can be found in the supplementary materials.

Our guided image SR method based on coupled dictionary learning falls into the learning-based category, as 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. Our approach also competes very well, especially in the presence of noise.

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 [34], [39]–[45]. However, these approaches differ from our proposed approach in a number of ways. For example, semi-coupled dictionary learning [34], supervised coupled dictionary learning [39], semi-supervised coupled dictionary learning [40], and semi-coupled low-rank discriminant dictionary learning [41] 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. [42] 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. [43] 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. The data model used in the proposed multimodal image SR approach is inspired from the data model proposed in [44], [45] 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 significantly different from [44], [45]. For example, we adapted K-SVD for coupled dictionary learning, developed multi-stage/cascaded model as an advanced version and incorporated neighbourhood regression for speed and performance improvement, etc. These designs are never presented in [44], [45].

III. Multimodal Image SR via Joint Sparse Representations Induced by Coupled Dictionaries

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

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 dependency 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:

$$\begin{aligned} \mathbf{x} &= \mathbf{\Psi}_c\,\mathbf{z} + \mathbf{\Psi}\,\mathbf{u}\,, \\ \mathbf{y} &= \mathbf{\Phi}_c\,\mathbf{z} + \mathbf{\Phi}\,\mathbf{v}\,, \end{aligned} \tag{1}$$

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, $\mathbf{\Psi}_c \in \mathbb{R}^{N_x \times K_c}$ and $\mathbf{\Phi}_c \in \mathbb{R}^{N_y \times K_c}$ are a pair of dictionaries associated with the common sparse representation $\mathbf{z}$, whereas $\mathbf{\Psi} \in \mathbb{R}^{N_x \times K_u}$ and $\mathbf{\Phi} \in \mathbb{R}^{N_y \times K_v}$ 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 the 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 target LR and guidance HR patches from different image modalities can be captured using sparse representations.

2. Second, we also assume – as in [8], [10], [33] – that the LR and HR patches from the same 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 the same image modality, and the guidance HR patch of another different image modality $\mathbf{y} \in \mathbb{R}^N$ as follows:[1]

$$\mathbf{x}^h = \mathbf{\Psi}_c^h\,\mathbf{z} + \mathbf{\Psi}^h\,\mathbf{u}\,, \tag{2}$$

$$\mathbf{x}^l = \mathbf{\Psi}_c^l\,\mathbf{z} + \mathbf{\Psi}^l\,\mathbf{u}\,, \tag{3}$$

$$\mathbf{y} = \mathbf{\Phi}_c\,\mathbf{z} + \mathbf{\Phi}\,\mathbf{v}\,, \tag{4}$$

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, $\mathbf{\Psi}_c^h \in \mathbb{R}^{N \times K}$, $\mathbf{\Psi}_c^l \in \mathbb{R}^{M \times K}$ and $\mathbf{\Phi}_c \in \mathbb{R}^{N \times K}$ are the dictionaries associated with the common sparse representation $\mathbf{z}$, whereas $\mathbf{\Psi}^h \in \mathbb{R}^{N \times K}$, $\mathbf{\Psi}^l \in \mathbb{R}^{M \times K}$ and $\mathbf{\Phi} \in \mathbb{R}^{N \times K}$ 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.

By capitalizing on this model, we propose in the sequel a novel guided image SR scheme that consists of two stages: a training stage referred to as coupled dictionary learning (CDL) and 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.

[1] Our 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 [34], [39]–[45], 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.

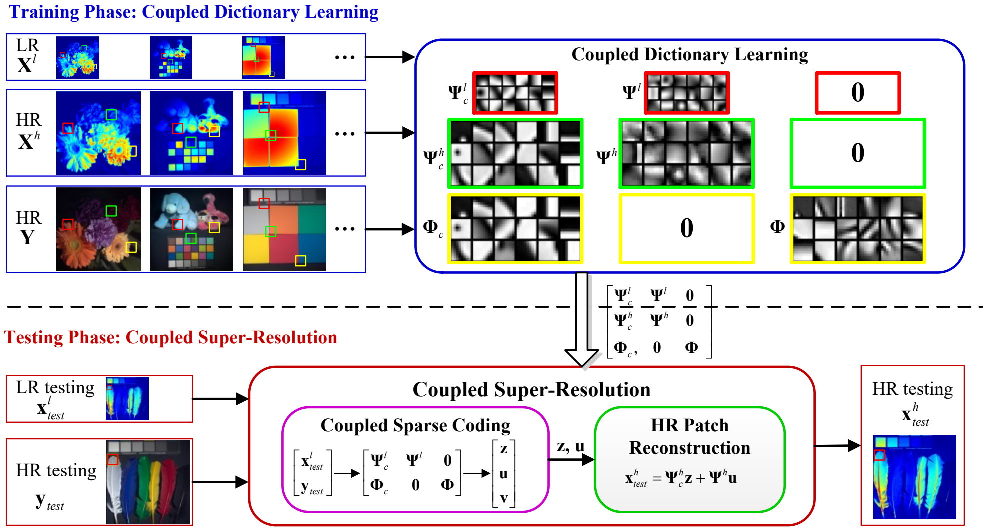

Training Phase: Coupled Dictionary Learning
LR $\mathbf{X}^l$
HR $\mathbf{X}^h$
HR $\mathbf{Y}$
Coupled Dictionary Learning
Ψ_c^l
Ψ^l
0
Ψ_c^h
Ψ^h
0
Φ_c
0
Φ
$\begin{bmatrix}\Psi_c^l & \Psi^l & 0\\ \Psi_c^h & \Psi^h & 0\\ \Phi_c & 0 & \Phi\end{bmatrix}$
Testing Phase: Coupled Super-Resolution
LR testing $\mathbf{x}^l_{test}$
HR testing $\mathbf{y}_{test}$
Coupled Super-Resolution
Coupled Sparse Coding
$\begin{bmatrix}\mathbf{x}^l_{test}\\ \mathbf{y}_{test}\end{bmatrix} \rightarrow \begin{bmatrix}\Psi_c^l & \Psi^l & 0\\ \Phi_c & 0 & \Phi\end{bmatrix} \rightarrow \begin{bmatrix}\mathbf{z}\\ \mathbf{u}\\ \mathbf{v}\end{bmatrix}$
z, u
HR Patch Reconstruction
$\mathbf{x}^h_{test} = \Psi^h_c \mathbf{z} + \Psi^h \mathbf{u}$
HR testing $\mathbf{x}^h_{test}$

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

Then, in the testing stage, we use the learned dictionaries to find the sparse 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 pairs of registered patches from 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 LR, HR, and the guidance image patches, respectively, and let $\mathbf{z}_i$, $\mathbf{u}_i$ and $\mathbf{v}_i$ ($i = 1 \dots T$) denote their sparse representations. The coupled dictionary learning problem can now be posed as follows:

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

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

Akin to other dictionary learning formulations [46], 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 $[\mathbf{\Psi}_c^l, \mathbf{\Psi}^l]$ and $[\mathbf{\Phi}_c, \mathbf{\Phi}]$ 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 $[\mathbf{\Psi}_c^h, \mathbf{\Psi}^h]$.[3] Algorithm 1 shows how we adapt K-SVD [46] accordingly.

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

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

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:

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{U},\mathbf{V}} \quad & \left\| \begin{bmatrix} \mathbf{X}^l\\ \mathbf{Y}\end{bmatrix} - \begin{bmatrix}\mathbf{\Psi}_c^l & \mathbf{\Psi}^l & \mathbf{0}\\ \mathbf{\Phi}_c & \mathbf{0} & \mathbf{\Phi}\end{bmatrix}\begin{bmatrix}\mathbf{Z}\\ \mathbf{U}\\ \mathbf{V}\end{bmatrix}\right\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{z}_i\|_0 + \|\mathbf{u}_i\|_0 + \|\mathbf{v}_i\|_0 \le s, \ \forall i.\end{aligned} \tag{7}$$

[2] Note that, we could also use alternative sparsity constraints, such as (a) $\|\mathbf{z}_i\|_0 + \|\mathbf{u}_i\|_0 \le s_x, \|\mathbf{z}_i\|_0 + \|\mathbf{v}_i\|_0 \le s_y$, (b) $\|\mathbf{z}_i\|_0 \le s_z, \|\mathbf{u}_i\|_0 \le s_u, \|\mathbf{v}_i\|_0 \le s_v$. 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.

[3] The 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.

Algorithm 1 Coupled Dictionary Learning

```
Input: Training data matrices $\mathbf{X}^l$, $\mathbf{X}^h$ and $\mathbf{Y}$.
Output: Dictionary pairs $[\boldsymbol{\Psi}_c^l, \boldsymbol{\Psi}^l]$, $[\boldsymbol{\Psi}_c^h, \boldsymbol{\Psi}^h]$ and $[\boldsymbol{\Phi}_c, \boldsymbol{\Phi}]$.
Initialization: Initialize dictionary atoms with randomly selected
    patches. Set the training iterations $OutIter$ and $InIter$, sparsity
    constraint $s$ and residual constraint $\epsilon$.
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:
                $(\hat{k}, \hat{\boldsymbol{\alpha}}) \in \underset{k\in\Gamma^c, \boldsymbol{\alpha}\in\mathbb{R}^{|\Gamma|+1}}{\arg\min} \left\| \begin{bmatrix} \mathbf{x}_i^l \\ \mathbf{y}_i \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Psi}_c^l & \boldsymbol{\Psi}^l & \mathbf{0} \\ \boldsymbol{\Phi}_c & \mathbf{0} & \boldsymbol{\Phi}_c \end{bmatrix}_{\Gamma\cup\{k\}} \boldsymbol{\alpha} \right\|_2^2$
                $\Gamma \leftarrow \Gamma \cup \{\hat{k}\}$; $[\mathbf{z}_i^T; \mathbf{u}_i^T; \mathbf{v}_i^T]_\Gamma \leftarrow \hat{\boldsymbol{\alpha}}$; $[\mathbf{z}_i^T; \mathbf{u}_i^T; \mathbf{v}_i^T]_{\Gamma^c} \leftarrow \mathbf{0}$
 8:         end while
 9:         Local Common Dictionary Update. Fix $\boldsymbol{\Psi}^l$, $\boldsymbol{\Phi}$, and only
            update $\boldsymbol{\Psi}_c^l$ and $\boldsymbol{\Phi}_c$ by solving (9). Specifically, for each
            atom pair $\begin{bmatrix} \boldsymbol{\psi}_{ck}^l \\ \boldsymbol{\phi}_{ck} \end{bmatrix}$ of $\begin{bmatrix} \boldsymbol{\Psi}_c^l \\ \boldsymbol{\Phi}_c \end{bmatrix}$, 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 that use $k$-th atom pair. Then,
            compute the representation residual
            $\mathbf{E}_k = \left( \begin{bmatrix} \mathbf{X}^l - \boldsymbol{\Psi}^l\mathbf{U} \\ \mathbf{Y} - \boldsymbol{\Phi}\mathbf{V} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Psi}_c^l \\ \boldsymbol{\Phi}_c \end{bmatrix} \mathbf{Z} + \begin{bmatrix} \boldsymbol{\psi}_{ck}^l \\ \boldsymbol{\phi}_{ck} \end{bmatrix} \mathbf{z}^k \right)_{(:,\Omega_k)}$
            Apply SVD on $\mathbf{E}_k = \mathbf{P}\boldsymbol{\Sigma}\mathbf{Q}^{\mathrm{T}}$ and choose the first column
            of $\mathbf{P}$ as the updated atom pair $\begin{bmatrix} \boldsymbol{\psi}_{ck}^l \\ \boldsymbol{\phi}_{ck} \end{bmatrix}$.
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 $\boldsymbol{\Psi}_c^l$, $\boldsymbol{\Phi}_c$, and only
            update $\boldsymbol{\Psi}^l$ and $\boldsymbol{\Phi}$ by solving (10) and (11). For each atom
            $\boldsymbol{\psi}_k^l$ of $\boldsymbol{\Psi}^l$, 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
            $\mathbf{E}_k = \left( \left[ \mathbf{X}^l - \boldsymbol{\Psi}_c^l \mathbf{Z} \right] - \boldsymbol{\Psi}^l \mathbf{U} + \boldsymbol{\psi}_k^l \mathbf{u}^k \right)_{(:,\Omega_k)}$
            Apply SVD on $\mathbf{E}_k = \mathbf{P}\boldsymbol{\Sigma}\mathbf{Q}^{\mathrm{T}}$ and choose the first column
            of $\mathbf{P}$ as the updated atom $\boldsymbol{\psi}_k^l$. Each atom $\boldsymbol{\phi}_k$ of $\boldsymbol{\Phi}$ is
            updated with $\Omega_k = \{i | 1 \leq i \leq T, \mathbf{v}^k(i) \neq 0\}$ and
            $\mathbf{E}_k = \left( \left[ \mathbf{Y} - \boldsymbol{\Phi}_c \mathbf{Z} \right] - \boldsymbol{\Phi}\mathbf{V} + \boldsymbol{\phi}_k \mathbf{v}^k \right)_{(:,\Omega_k)}$ in a similar
            manner.
14:     end for
15: end for
16: Step 2 – HR Dictionary learning:
17: Construct $[\boldsymbol{\Psi}_c^h, \boldsymbol{\Psi}^h]$ as in (13).
18: Return dictionaries.
```

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 [47].[4]

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

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

To this end, we adapt the K-SVD [46] 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. We first fix the unique dictionaries $\boldsymbol{\Psi}^l$, $\boldsymbol{\Phi}$ and only update the common dictionaries $\boldsymbol{\Psi}_c^l$ and $\boldsymbol{\Phi}_c$ by solving

$$\min_{\boldsymbol{\Psi}_c^l, \boldsymbol{\Phi}_c} \quad \left\| \begin{bmatrix} \mathbf{X}^l - \boldsymbol{\Psi}^l\mathbf{U} \\ \mathbf{Y} - \boldsymbol{\Phi}\mathbf{V} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Psi}_c^l \\ \boldsymbol{\Phi}_c \end{bmatrix} \mathbf{Z} \right\|_F^2 . \tag{9}$$

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:

$$\min_{\boldsymbol{\Psi}^l} \quad \left\| \left(\mathbf{X}^l - \boldsymbol{\Psi}_c^l \mathbf{Z} \right) - \boldsymbol{\Psi}^l \mathbf{U} \right\|_F^2 . \tag{10}$$

$$\min_{\boldsymbol{\Phi}} \quad \left\| \left(\mathbf{Y} - \boldsymbol{\Phi}_c \mathbf{Z} \right) - \boldsymbol{\Phi} \mathbf{V} \right\|_F^2 . \tag{11}$$

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

$$\min_{\boldsymbol{\Psi}_c^h, \boldsymbol{\Psi}^h} \left\| \mathbf{X}^h - \boldsymbol{\Psi}_c^h \mathbf{Z} - \boldsymbol{\Psi}^h \mathbf{U} \right\|_F^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{\Psi}_c^h & \boldsymbol{\Psi}^h \end{bmatrix} \right\|_F^2 \tag{12}$$

where the second term serves as a regularizer that makes the solution more stable[6]. This optimization problem – which exploits the conventional sparse representation invariance assumption that HR image patches $\mathbf{X}^h$ share the same sparse

[4] An 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.

[5] Owing to the SVD operation in the dictionary update, atoms from the common dictionary pair $[\boldsymbol{\Psi}_c^l; \boldsymbol{\Phi}_c]$ and the unique dictionaries $\boldsymbol{\Psi}^l$ and $\boldsymbol{\Phi}$ have unit ℓ_2 norm automatically.

[6] In 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 [48]. However, when there are rows of zeros or near-zeros 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).

Algorithm 2 Coupled Super-resolution

```
Input: The testing patch $\mathbf{x}^l_{test}$ and side information $\mathbf{y}_{test}$.
    Learned dictionaries $[\mathbf{\Psi}^l_c, \mathbf{\Psi}^l]$, $[\mathbf{\Psi}^h_c, \mathbf{\Psi}^h]$ and $[\mathbf{\Phi}_c, \mathbf{\Phi}]$.
Output: High resolution estimation $\mathbf{x}^h_{test}$.
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 as in (15).
```

codes with the corresponding LR version $\mathbf{X}^l$ – admits the closed form solution

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

Similar to conventional dictionary learning, our CDL algorithm cannot guarantee the convergence to a global optimum due to the non-convex nature of Problem (5). However, the CDL problem is convex with respect to the dictionaries when the sparse codes are fixed. When the dictionaries are fixed, it leads to a non-convex sparse coding problem with ℓ_0 pseudo-norm as constraints, but this can be solved efficiently using greedy algorithms. Therefore, the alternating optimization manner is adopted to address the CDL problem as in classical dictionary learning.

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 $\mathbf{x}^l_{test} \in \mathbb{R}^M$ denote a LR testing image patch and let $\mathbf{y}^h_{test} \in \mathbb{R}^N$ 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

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

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 [47] and iterative hard-thresholding algorithm [49] – 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 formulation (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 [21], [22], one can expect to obtain more accurate estimation of the sparse codes.

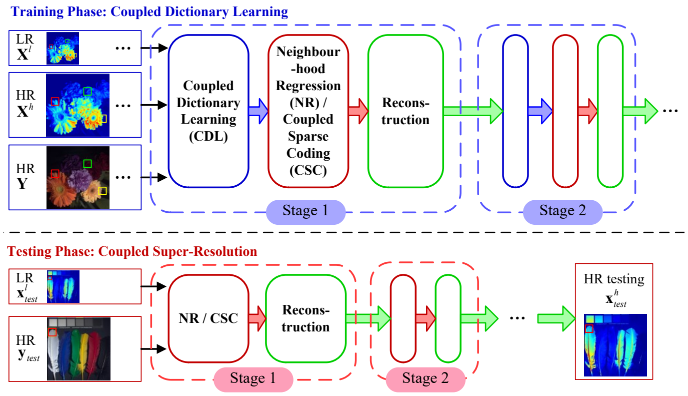

Training Phase: Coupled Dictionary Learning
LR $\mathbf{X}^l$
HR $\mathbf{X}^h$
HR $\mathbf{Y}$
Coupled Dictionary Learning (CDL)
Neighbour -hood Regression (NR) / Coupled Sparse Coding (CSC)
Recons- truction
Stage 1
Stage 2
Testing Phase: Coupled Super-Resolution
LR $\mathbf{x}^l_{test}$
HR $\mathbf{y}_{test}$
NR / CSC
Recons- truction
Stage 1
Stage 2
HR testing $\mathbf{x}^h_{test}$

Figure 2: Proposed multi-stage multi-modal image super-resolution approach with optional neighbourhood regression. $\mathbf{X}$ (or $\mathbf{x}$) and $\mathbf{Y}$ (or $\mathbf{y}$) represent the target and guidance modalities, respectively.

2) Step 2 – HR Patch Reconstruction: Finally, we can obtain an estimated HR patch of the target image $\mathbf{x}^h_{test}$ from the HR dictionaries $[\mathbf{\Psi}^h_c, \mathbf{\Psi}^h]$ and sparse codes $\mathbf{z}$ and $\mathbf{u}$ via:

$$\mathbf{x}^h_{test} = \mathbf{\Psi}^h_c\mathbf{z} + \mathbf{\Psi}^h\mathbf{u}\,. \quad (15)$$

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. Multi-Stage Multi-Modal Image Super-Resolution

The aforementioned coupled dictionary learning and coupled super-resolution constitute the basic version of our approach. In order to further exploit the power of the proposed model, we now introduce the multi-stage version, referred to as multi-stage CDLSR, consisting of multiple CDL and CSR stages. As shown in Figure 2, the CDL and CSR in each stage is the same as in the basic version. However, given the output of stage j, we perform extra CDL and CSR operations in stage $j + 1$, where the coupled dictionaries are trained using the estimated HR images of stage j in order to capture the new mapping to the groundtruth. In this way, the estimation in stage $j + 1$ is better than the estimation in stage j. This multi-stage strategy, also called stage-wise, hierarchical, or cascaded learning strategy, has been used in other works, such as the weighted analysis sparse representation (WASR) model [28], or the cascaded sparse coding network [18]. The difference between our multi-stage strategy and such works is that in each stage, we perform coupled dictionary learning, neighbourhood regression or coupled sparse coding so as to capture the dependencies among multimodal data in learned sparse domains. These multi-stage CDL and CSR operations along with neighbourhood regression are further described below.

A. Multi-stage CDL

The coupled dictionary learning problem in the multi-stage framework at stage j can be formulated as:

$$\begin{aligned} \underset{\mathbf{D}^{(j)},\boldsymbol{\alpha}^{(j)}}{\text{minimize}} \ & \left\| \begin{bmatrix} \mathbf{X}^{l^{(j)}} \\ \mathbf{X}^{h} \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{\Psi}_c^{l^{(j)}} & \mathbf{\Psi}^{l^{(j)}} & \mathbf{0} \\ \mathbf{\Psi}_c^{h^{(j)}} & \mathbf{\Psi}^{h^{(j)}} & \mathbf{0} \\ \mathbf{\Phi}_c{}^{(j)} & \mathbf{0} & \mathbf{\Phi}^{(j)} \end{bmatrix} \begin{bmatrix} \mathbf{Z}^{(j)} \\ \mathbf{U}^{(j)} \\ \mathbf{V}^{(j)} \end{bmatrix} \right\|_F^2 \\ \text{subject to} \ & \|\mathbf{z}_i{}^{(j)}\|_0 + \|\mathbf{u}_i{}^{(j)}\|_0 + \|\mathbf{v}_i{}^{(j)}\|_0 \leq s, \ \forall i, \end{aligned} \tag{16}$$

where $\mathbf{X}^h$ and $\mathbf{Y}$ denote HR target and guidance modalities, the same as in (5). $\mathbf{X}^{l^{(j)}}$ denotes the LR input in the j-th stage. Specifically, for the first stage $j = 1$, $\mathbf{X}^{l^{(j)}}$ denotes the original LR input, and for the second and subsequent stages $j \geq 2$, $\mathbf{X}^{l^{(j)}}$ denotes the estimated high-resolution output using (17) and (18) from the previous stage, that is, $\mathbf{X}^{l^{(j)}} = \mathbf{X}^{h^{(j-1)}}$. $\mathbf{D}^{(j)}$ and $\boldsymbol{\alpha}^{(j)}$ denotes the set of coupled dictionaries $\{\mathbf{\Psi}_c^{l^{(j)}}, \mathbf{\Psi}^{l^{(j)}}, \mathbf{\Psi}_c^{h^{(j)}}, \mathbf{\Psi}^{h^{(j)}}, \mathbf{\Phi}_c{}^{(j)}, \mathbf{\Phi}^{(j)}\}$ and sparse representations $\{\mathbf{Z}^{(j)}, \mathbf{U}^{(j)}, \mathbf{V}^{(j)}\}$ learned in the j-th stage. Algorithm 1 is still available for coupled dictionary learning in each stage.

B. Multi-stage CSR

Given the learned coupled dictionaries associated with each stage, the multi-stage coupled super-resolution problem at stage j can be formulated as following (17) and (18):

$$\begin{aligned} \min_{\boldsymbol{\alpha}^{(j)}} \ & \left\| \begin{bmatrix} \mathbf{x}^{l^{(j)}}_{test} \\ \mathbf{y}_{test} \end{bmatrix} - \begin{bmatrix} \mathbf{\Psi}_c^{l^{(j)}} & \mathbf{\Psi}^{l^{(j)}} & \mathbf{0} \\ \mathbf{\Phi}_c{}^{(j)} & \mathbf{0} & \mathbf{\Phi}^{(j)} \end{bmatrix} \begin{bmatrix} \mathbf{z}^{(j)} \\ \mathbf{u}^{(j)} \\ \mathbf{v}^{(j)} \end{bmatrix} \right\|_F^2 \\ \text{s.t.} \ & \|\mathbf{z}^{(j)}\|_0 + \|\mathbf{u}^{(j)}\|_0 + \|\mathbf{v}^{(j)}\|_0 \leq s \,, \end{aligned} \tag{17}$$

$$\mathbf{x}^{h^{(j)}}_{test} = \mathbf{\Psi}_c^{h^{(j)}} \mathbf{z}^{(j)} + \mathbf{\Psi}^{h^{(j)}} \mathbf{u}^{(j)} \,. \tag{18}$$

Once all the HR patches are recovered, they are integrated into a whole image by averaging on the overlapping areas. The estimation in the j-th stage is then input to the next stage as a new LR image.

C. Multi-stage CSR with Neighbourhood Regression

Instead of performing time-consuming coupled sparse coding as in Section III-C, we can also integrate neighbourhood regression into our framework to accelerate the inference speed and increase the performance. Specifically, taking the stage j for example (we omit the superscript $^{(j)}$ for notation simplicity), we use each atom pair $[\psi^l_{ck}; \phi_{ck}]$ from the learned coupled dictionaries $[\mathbf{\Psi}^l_c; \mathbf{\Phi}_c]$ as a centroid (or called anchor) to construct its neighbourhood using its K nearest patch pairs $\mathbf{N}_l = [\mathbf{x}^l_k; \mathbf{y}_k] (k = 1...K)$. These K normalized patch pairs together with the centroid constitute a new set of coupled sub-dictionaries $\mathbf{N}_l$ (or called regressors). A similar operation is also applied to each atom ψ^l_k from $\mathbf{\Psi}^l$ and each atom ϕ_k from $\mathbf{\Phi}$. During the testing phase, for each testing patch pair, we find its closest centroid and corresponding sub-dictionaries $\mathbf{N}_l$ to perform ridge regression in this neighbourhood in order to get the sparse codes $\mathbf{z}$ and $\mathbf{u}$, as shown in (19), which together with (15) lead to a closed formulation for the reconstruction of the high-resolution patch as (20).

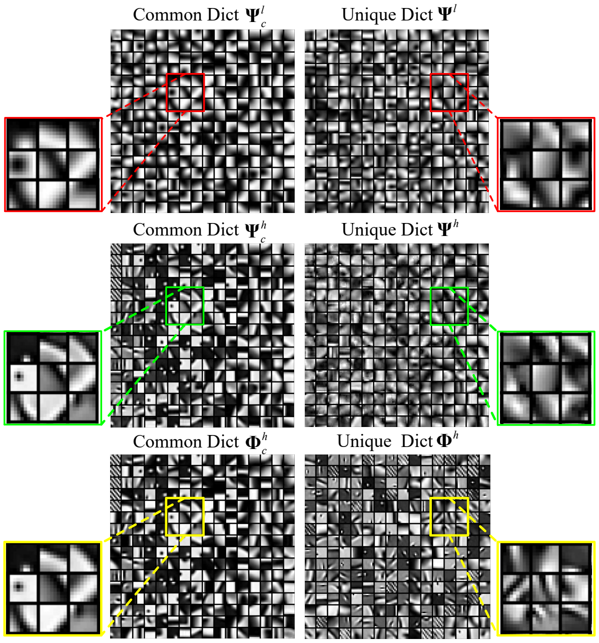

Common Dict $\mathbf{\Psi}_c^l$
Unique Dict $\mathbf{\Psi}^l$
Common Dict $\mathbf{\Psi}_c^h$
Unique Dict $\mathbf{\Psi}^h$
Common Dict $\mathbf{\Phi}_c^h$
Unique Dict $\mathbf{\Phi}^h$

Figure 3: Learned coupled dictionaries for multi-spectral images of wavelength 640nm and RGB images using the basic version of our algorithm. 256 atoms are shown here. The first row indicates the common and unique dictionaries learned from 4× downsampling LR multi-spectral images. The second row indicates the HR dictionary pair. The last row shows the dictionaries learned from guidance RGB modality.

$$\min_{\boldsymbol{\alpha}^{(j)}} \ \left\| \begin{bmatrix} \mathbf{x}^{l^{(j)}}_{test} \\ \mathbf{y}_{test} \end{bmatrix} - \mathbf{N}_l{}^{(j)} \boldsymbol{\alpha}^{(j)} \right\|_2^2 + \gamma \left\| \boldsymbol{\alpha}^{(j)} \right\|_2^2 \ , \tag{19}$$

$$\mathbf{x}^{h^{(j)}}_{test} = \mathbf{N}_h{}^{(j)} (\mathbf{N}_l{}^{(j)\top} \mathbf{N}_l{}^{(j)} + \gamma \mathbf{I})^{-1} \mathbf{N}_l{}^{(j)\top} \begin{bmatrix} \mathbf{x}^{l^{(j)}}_{test} \\ \mathbf{y}_{test} \end{bmatrix} , \tag{20}$$

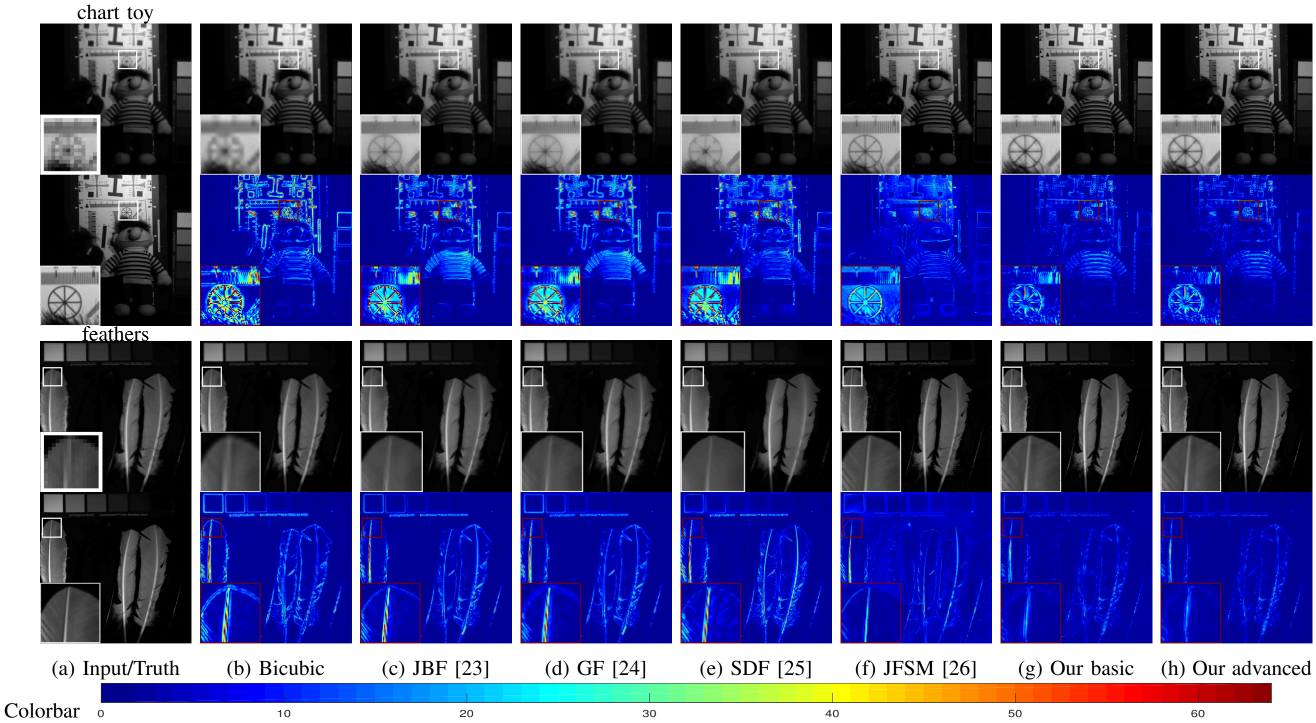
where $\mathbf{N}_l{}^{(j)}$ denotes the sub-dictionary of LR target and HR guidance modality in the selected neighbourhood and $\mathbf{N}_h{}^{(j)}$ denotes the sub-dictionary of HR target modality.

The neighbourhood regression concept is modified from the adjusted anchored neighbourhood regression [13] to work with multimodal data and coupled dictionaries. As each patch usually lies in a local linear low-dimensional manifold, it can be well approximated using a linear combination of sub-dictionaries in its neighbourhood. Owing to this property, we can take better advantage of a large amount of training samples to generate high-quality neighbourhood for the coupled dictionaries, thereby improving the representation performance.

V. Experiments

We now present a series of experiments to validate the effectiveness of the proposed joint image SR approach in various scenarios.

In subsection V-A, we compare our approaches with representative approaches that exploit "shallow" models for multimodal image SR, including the Joint Bilateral Filtering (JBF) [23], the Guided image Filtering (GF) [24], the Static/Dynamic Filtering (SDF) [25], and the Joint Filtering via optimizing a Scale Map (JFSM) [26].



(a) Input/Truth (b) Bicubic (c) JBF [23] (d) GF [24] (e) SDF [25] (f) JFSM [26] (g) Our basic (h) Our advanced

Figure 4: Comparison with "shallow" models on 4× upscaling for multi-spectral images of 640nm band. 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: Comparison with "shallow" models on 4× and 6× upscaling for multi-spectral image of 640 nm band.

4 ×	Bicubic		JBF [23]		GF [24]		SDF [25]		JFSM [26]		Our basic		Our advanced	
	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	0.9215	33.30	0.9855	34.50	**0.9875**	**35.79**
cloth	0.7571	26.91	0.7640	27.62	0.7699	27.79	0.7315	27.18	**0.9770**	**35.33**	0.9506	32.75	0.9493	33.68
egyptian	0.9761	36.22	0.9788	37.82	0.9788	37.96	0.9677	37.16	0.9428	39.68	0.9935	**42.63**	**0.9938**	42.62
feathers	0.9530	30.46	0.9599	31.80	0.9618	32.12	0.9434	30.92	0.9096	33.54	0.9871	36.25	**0.9887**	**37.25**
glass tiles	0.9215	26.38	0.9339	27.15	0.9326	27.45	0.9188	27.01	0.9407	29.34	0.9791	31.05	**0.9828**	**32.26**
jelly beans	0.9269	27.45	0.9474	28.97	0.9488	29.54	0.9279	27.87	0.9356	30.82	0.9866	34.38	**0.9886**	**35.72**
oil painting	0.9025	32.23	0.9034	33.23	0.9033	33.30	0.9001	32.80	0.9439	34.16	**0.9601**	36.24	0.9556	**36.27**
paints	0.9569	30.47	0.9714	32.08	0.9698	32.23	0.9569	31.35	0.9321	32.96	0.9900	36.99	**0.9907**	**37.24**
average	0.9174	29.91	0.9265	31.17	0.9270	31.39	0.9123	30.63	0.9379	33.64	0.9791	35.60	**0.9796**	**36.36**

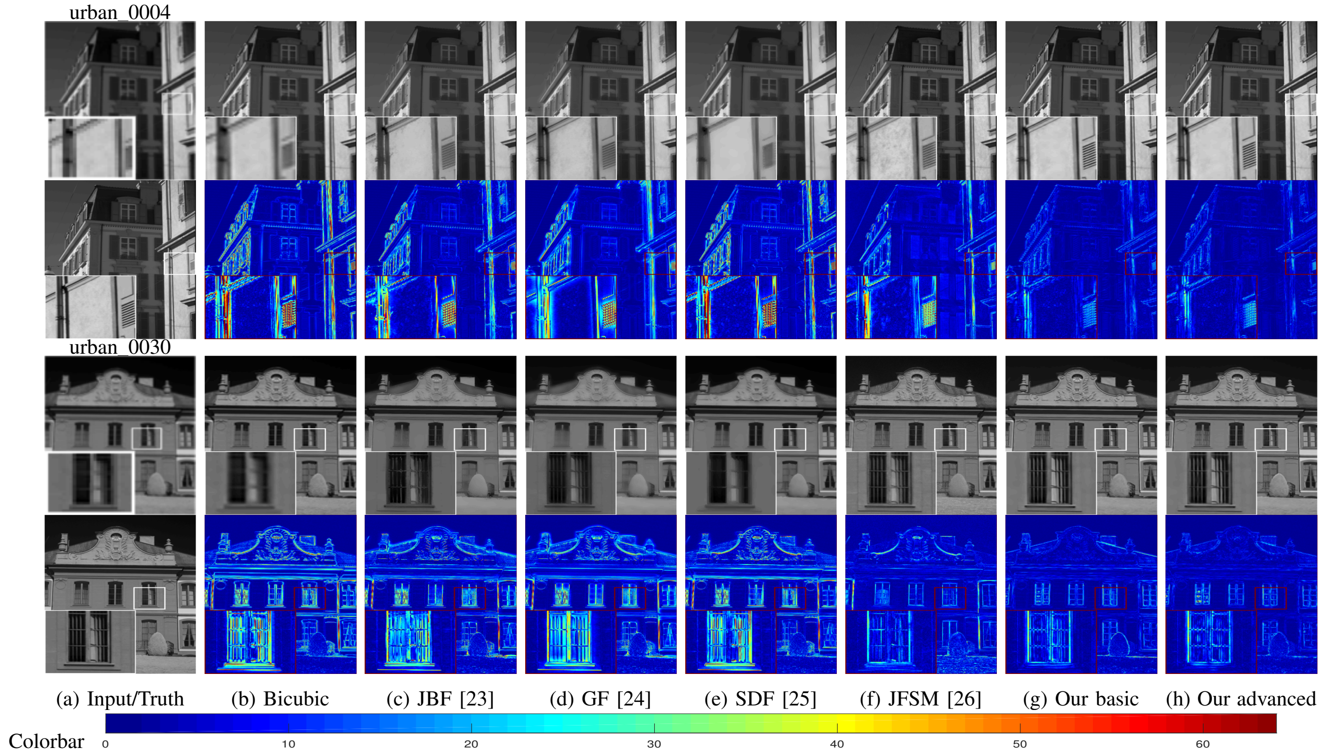
6 ×	Bicubic		JBF [23]		GF [24]		SDF [25]		JFSM [26]		Our basic		Our advanced	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
chart toy	0.8992	28.08	0.8992	28.08	0.8932	27.86	0.9006	28.12	0.9144	31.22	**0.9682**	32.55	0.9675	**32.58**
cloth	0.6424	26.06	0.6424	26.06	0.6394	26.07	0.6158	25.80	**0.9723**	**33.79**	0.9256	31.73	0.9163	31.50
egyptian	0.9560	34.95	0.9560	34.95	0.9536	34.80	0.9466	34.83	0.9444	38.43	**0.9872**	**40.75**	0.9860	40.42
feathers	0.9177	28.80	0.9177	28.80	0.9138	28.76	0.9062	28.50	0.9042	31.32	0.9727	33.75	**0.9763**	**34.29**
glass tiles	0.8652	25.05	0.8652	25.05	0.8585	25.06	0.8556	25.03	0.9233	27.33	0.9646	29.87	**0.9705**	**30.76**
jelly beans	0.8835	26.24	0.8835	26.24	0.8801	26.36	0.8681	25.60	0.9225	28.58	0.9734	32.73	**0.9754**	**32.92**
oil painting	0.8664	31.87	0.8664	31.87	0.8626	31.78	0.8574	31.49	0.9462	34.09	0.9427	35.18	**0.9517**	**35.98**
paints	0.9328	29.04	0.9328	29.04	0.9253	28.90	0.9226	28.69	0.9363	31.25	0.9792	34.93	**0.9820**	**35.91**
average	0.8704	28.76	0.8704	28.76	0.8658	28.70	0.8591	28.50	0.9330	32.00	0.9642	33.93	**0.9657**	**34.29**

In subsection V-B, we compare our approaches with state-of-the-art approaches that exploit "deep" models for multi-modal image SR, including the Deep Joint image Filtering (DJF) [27] and the weighted analysis sparse representation model (WASR) [28], as well as popular single-modal image SR methods, including Super-Resolution Convolutional Neural Network (SRCNN) [14], Fast Super-Resolution Convolutional Neural Network (FSRCNN) [15], and Cascaded Sparse Coding Network (CSC-Net) [18].

We adopt the Peak Signal to Noise Ratio (PSNR), the root-mean-square error (RMSE) and the Structure SIMilarity (SSIM) index [50] 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 database[7]. The infrared/RGB images datasets are obtained from the EPFL RGB-NIR Scene database[8]. All these datasets are registered for both modalities. For each multimodal dataset, we randomly separate the 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 × and 6 ×, using the MATLAB "imresize" function to generate corresponding LR versions, similar to

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

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



(a) Input/Truth (b) Bicubic (c) JBF [23] (d) GF [24] (e) SDF [25] (f) JFSM [26] (g) Our basic (h) Our advanced

Figure 5: Comparison with "shallow" models on 4× upscaling for near-infrared house images. 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 II: Comparison with "shallow" models on 4× and 6× upscaling for near-infrared house images.

4×	Bicubic		JBF [23]		GF [24]		SDF [25]		JFSM [26]		Our basic		Our advanced	
	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	0.9721	30.86	0.9811	34.14	**0.9828**	**34.70**
urban_0006	0.9458	30.89	0.9311	32.10	0.9400	32.66	0.8918	30.60	0.9741	32.86	0.9868	36.79	**0.9889**	**37.71**
urban_0017	0.9527	30.45	0.9172	31.11	0.9205	31.32	0.9281	30.72	0.9500	32.85	0.9777	35.27	**0.9830**	**36.05**
urban_0018	0.9298	25.19	0.9308	27.59	0.9251	27.70	0.9196	26.09	0.9774	30.80	0.9874	33.01	**0.9895**	**33.81**
urban_0020	0.9577	28.03	0.9523	30.67	0.9494	30.69	0.9505	29.09	0.9797	32.61	0.9893	36.66	**0.9919**	**37.83**
urban_0026	0.8704	26.27	0.8627	26.82	0.8571	26.89	0.8558	26.61	0.9332	28.97	0.9482	30.35	**0.9548**	**30.91**
urban_0030	0.8401	26.54	0.8476	27.58	0.8383	27.59	0.8415	27.21	0.9064	30.56	0.9443	32.71	**0.9548**	**33.73**
urban_0050	0.9434	26.65	0.9099	27.32	0.9116	27.35	0.9207	27.07	0.9251	27.58	0.9663	29.37	**0.9706**	**29.69**
average	0.9179	27.49	0.9109	28.96	0.9101	29.12	0.9018	28.03	0.9522	30.89	0.9726	33.54	**0.9770**	**34.30**
6×	Bicubic		JBF [23]		GF [24]		SDF [25]		JFSM [26]		Our basic		Our advanced	
	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	0.9527	27.97	0.9558	30.77	**0.9660**	**32.79**
urban_0006	0.8671	28.48	0.8861	30.00	0.8876	30.17	0.8377	28.76	0.9716	32.28	0.9664	34.15	**0.9714**	**35.45**
urban_0017	0.8998	28.64	0.8864	29.63	0.8860	29.61	0.8910	29.13	0.9434	32.01	0.9515	32.98	**0.9620**	**34.23**
urban_0018	0.8393	23.07	0.8718	25.09	0.8591	24.98	0.8439	23.79	0.9470	27.47	0.9727	31.03	**0.9776**	**32.14**
urban_0020	0.9053	26.03	0.9200	28.19	0.9118	28.01	0.9089	26.93	0.9673	30.33	0.9763	33.85	**0.9819**	**35.82**
urban_0026	0.7850	24.71	0.8235	25.64	0.8131	25.63	0.7989	25.17	0.9128	27.54	0.9172	28.88	**0.9265**	**29.57**
urban_0030	0.7517	25.19	0.7994	26.32	0.7855	26.22	0.7748	25.80	0.8902	29.38	0.9099	30.52	**0.9230**	**31.74**
urban_0050	0.8921	25.17	0.8837	26.26	0.8846	26.26	0.8837	25.90	0.9068	26.67	0.9402	28.37	**0.9461**	**28.67**
average	0.8437	25.65	0.8696	27.13	0.8637	27.10	0.8475	26.26	0.9365	29.21	0.9487	31.32	**0.9568**	**32.55**

Table III: Comparison with "shallow" models on 4× upscaling for near-infrared landscape images.

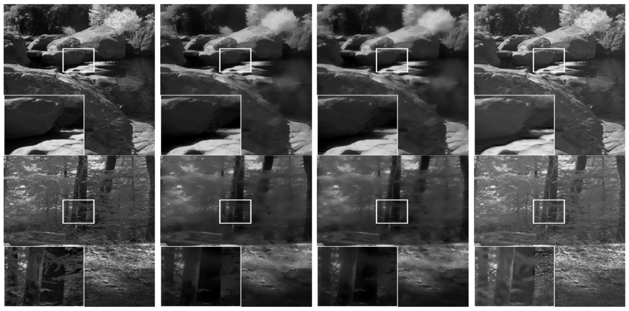
	JBF [23]		GF [24]		JFSM [26]		Our basic	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
n0025	0.8022	27.08	0.7970	27.14	0.8242	25.67	**0.9097**	**29.05**
n0027	0.6880	25.55	0.7002	25.82	0.7033	24.68	**0.8702**	**28.07**
n0028	0.7358	24.82	0.7519	25.01	0.7766	24.16	**0.8789**	**26.50**
n0031	0.8452	27.58	0.8524	27.81	0.8536	26.71	**0.9136**	**28.64**
n0049	0.7720	29.20	0.7832	29.52	0.7453	26.85	**0.8996**	**31.88**
n0051	0.7287	26.01	0.7262	25.97	0.7606	25.19	**0.8767**	**28.29**
average	0.7620	26.71	0.7685	26.88	0.7773	25.54	**0.8914**	**28.74**

[8], [33]. These settings and procedures are applied to all the experiments.

A. *Comparison with "Shallow" Models*

In this sub-section, we compare the basic and advanced version of our model with representative approaches that exploit "shallow" models [23]–[26], using bicubic interpolation as the baseline method. The experiments include multi-spectral image super-resolution and near-infrared image super-resolution aided by their corresponding RGB modality.

Training Phase with CDL. We adopt some common pre-processing operations to prepare the training dataset. Specifi-



Truth JBF [23] GF [24] Our basic

Figure 6: Comparison with "shallow" models on 4× upscaling for near-infrared landscape images, e.g., n0031(up) and n0051(bottom).

cally, 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 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 $\sqrt{N} \times \sqrt{N}$ 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 $[\mathbf{\Psi}_c^l, \mathbf{\Psi}^l]$ and $[\mathbf{\Phi}_c, \mathbf{\Phi}]$ from $\mathbf{X}^l$ and $\mathbf{Y}$. Then, HR dictionary pair $[\mathbf{\Psi}_c^h, \mathbf{\Psi}^h]$ 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× upscaling and 16×16 for 6× upscaling, dictionary size $K = 1024$, total sparsity constraint $s = 20$, training size $T \approx 15,000$.

Figure 3 shows the learned coupled dictionaries using the basic version of our algorithm 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 $\mathbf{\Psi}_c^l$ and $\mathbf{\Psi}_c^h$ capture associated edges, blobs, textures with the same direction and location. Similar behavior can also be observed in $\mathbf{\Psi}^l$ and $\mathbf{\Psi}^h$. 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 $\mathbf{\Phi}_c^h$ from the guidance images exhibits considerable resemblance and strong correlation to $\mathbf{\Psi}_c^h$ and $\mathbf{\Psi}_c^l$ from the target HR/LR modalities. This indicates that the three common dictionaries have indeed captured the similarities between multi-spectral and RGB modalities. In contrast, the learned unique dictionaries $\mathbf{\Psi}^h$ and $\mathbf{\Phi}$ 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 target image and HR guidance image for test, we upscale the LR target image to the desired size as before. Then the testing image pairs 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 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 $[\mathbf{\Psi}_c^l, \mathbf{\Psi}^l]$ and $[\mathbf{\Phi}_c, \mathbf{\Phi}]$ 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 $[\mathbf{\Psi}_c^h, \mathbf{\Psi}^h]$ 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.

Table IV: Comparison with "deep" models on 4× and 6× upscaling for near-infrared images.

4×	WASR [28]		DJF [27]		Our basic		Our advanced	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
u_0004	0.9786	32.25	0.9895	35.85	0.9811	34.14	**0.9896**	**35.87**
u_0006	0.9866	36.71	0.9917	37.92	0.9868	36.79	**0.9933**	**38.31**
u_0017	0.9799	35.11	0.9860	36.99	0.9777	35.27	**0.9889**	**37.33**
u_0018	0.9840	31.44	0.9932	34.44	0.9874	33.01	**0.9939**	**35.08**
u_0020	0.9892	35.33	0.9953	38.22	0.9893	36.66	**0.9959**	**39.36**
u_0026	0.9490	30.01	**0.9635**	**31.52**	0.9482	30.35	**0.9635**	31.46
u_0030	0.9346	31.10	**0.9601**	**35.07**	0.9443	32.71	0.9570	34.22
u_0050	0.9674	29.30	0.9589	28.34	0.9663	29.37	**0.9745**	**30.13**
average	0.9711	32.66	0.9798	34.79	0.9726	33.54	**0.9821**	**35.22**

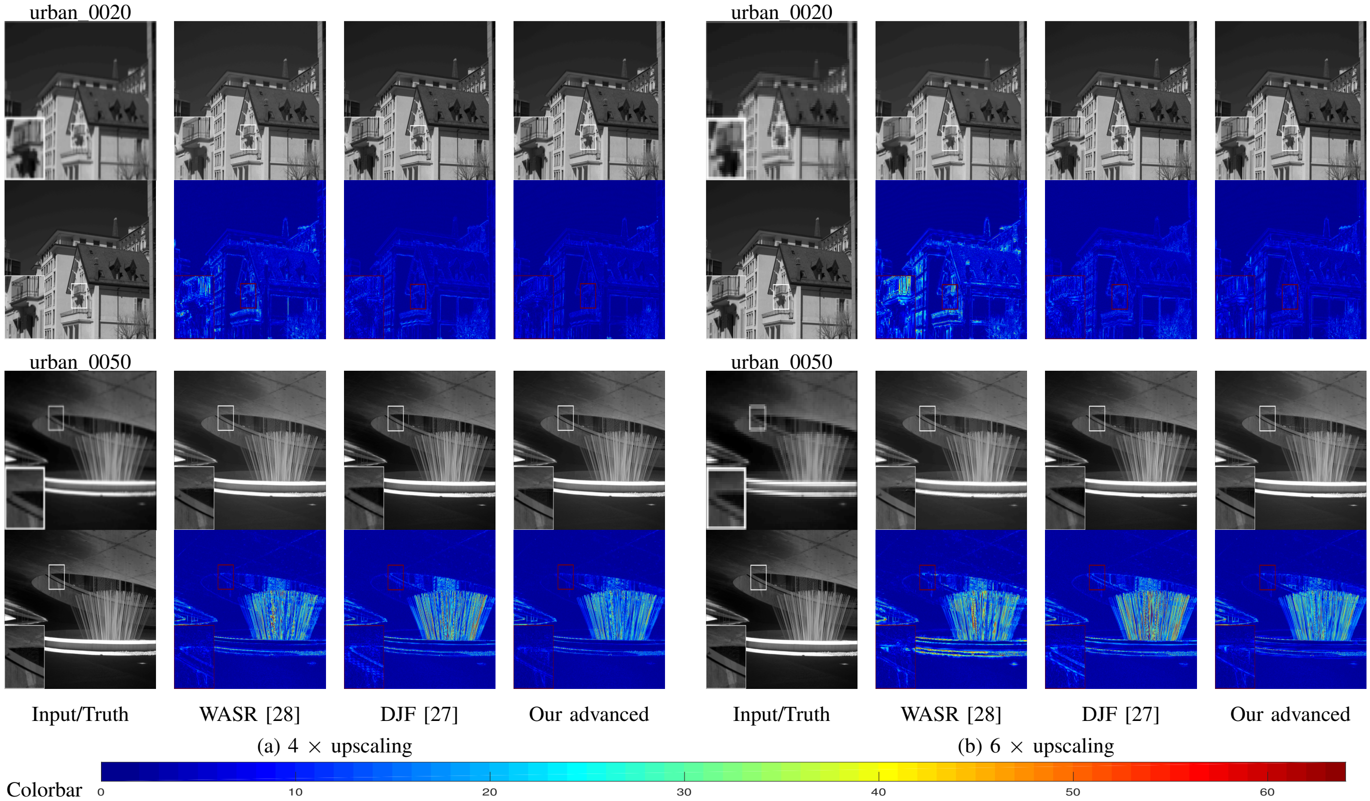
6×	WASR [28]		DJF [27]		Our basic		Our advanced	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
u_0004	0.9466	29.03	**0.9814**	33.85	0.9558	30.77	0.9805	**34.05**
u_0006	0.9702	33.96	0.9851	**36.19**	0.9664	34.15	**0.9859**	**36.19**
u_0017	0.9517	32.55	0.9713	**35.28**	0.9515	32.98	**0.9748**	35.26
u_0018	0.9565	28.80	0.9880	33.05	0.9727	31.03	**0.9886**	**33.50**
u_0020	0.9726	32.50	0.9908	36.88	0.9763	33.85	**0.9911**	**37.34**
u_0026	0.9069	28.32	**0.9421**	**30.19**	0.9172	28.88	0.9380	29.97
u_0030	0.8812	28.26	**0.9365**	**33.34**	0.9099	30.52	0.9245	31.90
u_0050	0.9375	27.28	0.9318	27.12	0.9402	28.37	**0.9547**	**29.06**
average	0.9404	30.09	0.9659	33.24	0.9487	31.32	**0.9673**	**33.41**

4×	SRCNN [14]		FSRCNN [15]		CSC-Net [18]		
single	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	
u_0004	0.9341	27.27	0.9371	27.42	0.9363	27.41	
u_0006	0.9672	32.31	0.9691	32.52	0.9679	32.36	
u_0017	0.9688	31.68	0.9701	31.76	0.9684	31.53	
u_0018	0.9618	26.92	0.9652	27.21	0.9630	27.00	
u_0020	0.9767	29.76	0.9777	29.84	0.9764	29.69	
u_0026	0.9130	27.61	0.9184	27.93	0.9145	27.71	
u_0030	0.8838	27.55	0.8852	27.67	0.8847	27.67	
u_0050	0.9626	27.95	0.9670	28.20	0.9660	28.20	
average	0.9460	28.88	0.9487	29.07	0.9472	28.95	

Multi-stage CDLSR. For the multi-stage version of the proposed algorithm, we perform aforementioned CDL and CSR operations in additional stages, shown in Figure 2. We utilize three stages in our experiments. Adding more stages may further improve the performance, but also suffers from high computational burden in both the training and testing phase. We also find that the three-stage CDLSR already brings significant improvements over our basic version, as well as the various other competing methods.

Figure 4 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 [23], GF [24], SDF [25] exhibit noticeable blurred areas. The reconstruction from JFSM [26] shows sharp edges but with weaker intensity than the ground-truth, a form of lu-

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



(a) 4 × upscaling

(b) 6 × upscaling

Figure 7: Comparison with "deep" models on 4× and 6× upscaling for near-infrared images.

minance 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. Moreover, the proposed advanced version further improves the performance. This is also confirmed by the error maps, as well as by quantitative measure in terms of PSNR and SSIM, shown in Table I for 4× and 6× upscaling, respectively. The quantitative results show that the basic version of our method 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× and 6× upscaling, the proposed approach outperforms JBF [23], GF [24], SDF [25], JFSM [26] with gains of at least 1.9dB in terms of average PSNR.

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 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 5 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. Moreover, our advanced version further improves the performance. Table II also confirms the significant advantage of the proposed approach over other competing 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 6 and Table III 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.

B. Comparison with "Deep" Models

In this subsection, we compare our approaches with state-of-the-art approaches that exploit "deep" models for multimodal or single-modal image SR [14], [15], [18], [27], [28]. We consider both the noise-free situation (there is no noise in the training and testing dataset), and the noisy situation (there existing noise in the training and/or testing dataset).

1) Noise-free Situation: We repeat the RGB guided NIR-SR experiments as in the Section V-A (comparison with "Shallow" models) with similar settings. We evaluate the performance of all the competing algorithms on a large training

Table V: Comparison with "deep" models on noisy LR testing images. There exists no training noise, i.e., $\sigma_{train} = 0$, but there exists testing noise with standard deviation $\sigma_{test} = [4, 8, 10, 12]$. $4\times$ upscaling for noisy near-infrared images with RGB modality for guidance.

	$\sigma_{train} = 0$, $\sigma_{test} = 4$								$\sigma_{train} = 0$, $\sigma_{test} = 8$							
	Our basic		Our advanced		WASR [28]		DJF [27]		Our basic		Our advanced		WASR [28]		DJF [27]	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.9715	32.96	0.9712	34.09	0.9576	31.68	0.9386	33.62	0.9477	31.97	0.9409	32.36	0.8972	29.95	0.8313	30.21
urban_0006	0.9752	35.77	0.9786	36.31	0.9665	35.07	0.9524	34.54	0.9483	33.71	0.9374	33.44	0.9116	32.01	0.8574	30.33
urban_0017	0.9592	34.62	0.9649	35.01	0.9539	33.97	0.9233	34.19	0.9243	33.08	0.9204	32.44	0.8797	31.53	0.7896	30.30
urban_0018	0.9801	32.58	0.9817	33.69	0.9643	30.86	0.9537	32.75	0.9594	31.88	0.9626	32.40	0.9090	29.34	0.8650	29.75
urban_0020	0.9784	36.11	0.9771	36.95	0.9575	34.06	0.9334	34.94	0.9479	34.51	0.9521	34.52	0.8720	31.36	0.8070	30.74
urban_0026	0.9339	30.00	0.9379	30.45	0.9270	29.58	0.9105	30.58	0.9104	29.46	0.9058	29.34	0.8661	28.39	0.7968	28.54
urban_0030	0.9278	32.25	0.9357	33.21	0.9054	30.50	0.9101	33.09	0.9043	31.37	0.9095	31.69	0.8368	29.15	0.8003	29.81
urban_0050	0.9465	28.92	0.9486	29.45	0.9288	28.86	0.8935	27.84	0.9138	28.54	0.9162	28.74	0.8254	27.66	0.7559	26.66
average	0.9591	32.90	**0.9620**	**33.64**	0.9451	31.82	0.9269	32.69	**0.9320**	31.82	0.9306	**31.87**	0.8747	29.92	0.8129	29.54

	$\sigma_{train} = 0$, $\sigma_{test} = 10$								$\sigma_{train} = 0$, $\sigma_{test} = 12$							
	Our basic		Our advanced		WASR [28]		DJF [27]		Our basic		Our advanced		WASR [28]		DJF [27]	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.9323	31.35	0.9268	31.54	0.8537	28.97	0.7745	28.72	0.9150	30.71	0.9139	30.75	0.8049	27.81	0.7181	27.42
urban_0006	0.9316	32.75	0.9128	32.34	0.8684	30.36	0.8031	28.64	0.9132	31.87	0.8881	31.42	0.8241	28.85	0.7493	27.23
urban_0017	0.9021	32.26	0.8953	31.49	0.8303	30.21	0.7166	28.71	0.8777	31.47	0.8743	30.71	0.7668	28.68	0.6571	27.33
urban_0018	0.9456	31.32	0.9495	31.59	0.8678	28.38	0.8172	28.34	0.9302	30.72	0.9380	30.82	0.8195	27.12	0.7766	27.13
urban_0020	0.9282	33.57	0.9396	33.55	0.8092	29.81	0.7406	29.10	0.9063	32.64	0.9270	32.64	0.7415	28.16	0.6807	27.63
urban_0026	0.8956	29.10	0.8860	28.71	0.8283	27.62	0.7427	27.52	0.8788	28.69	0.8650	28.11	0.7704	26.48	0.6808	26.44
urban_0030	0.8903	30.84	0.8972	30.97	0.7816	28.00	0.7442	28.41	0.8754	30.29	0.8860	30.37	0.7242	26.81	0.6907	27.06
urban_0050	0.8924	28.29	0.9004	28.38	0.7619	26.87	0.6874	25.95	0.8687	28.01	0.8858	28.03	0.6911	25.85	0.6270	25.23
average	**0.9148**	**31.18**	0.9135	31.07	0.8251	28.78	0.7533	28.17	0.8956	**30.55**	**0.8973**	30.36	0.7678	27.47	0.6975	26.93

Table VI: Comparison with "deep" models on noisy both LR testing and LR training images. There exists training noise with standard deviation $\sigma_{train} = 12$, and there also exists testing noise with standard deviation $\sigma_{test} = [12, 13.2, 14.4, 15.6]$, corresponding to mismatch $\delta = [0, 10\%, 20\%, 30\%]$. $4\times$ upscaling for noisy near-infrared images with RGB modality for guidance.

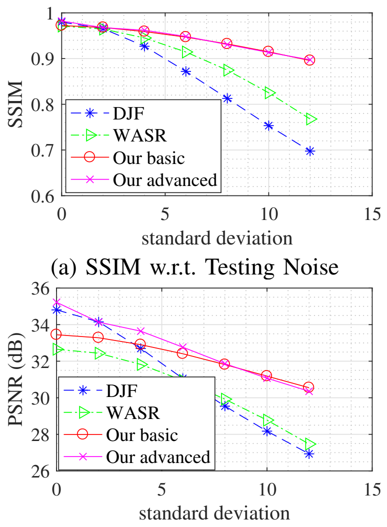
	$\sigma_{train} = 12$, $\sigma_{test} = 12$								$\sigma_{train} = 12$, $\sigma_{test} = 13.2$							
	Our basic		Our advanced		WASR [28]		DJF [27]		Our basic		Our advanced		WASR [28]		DJF [27]	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.9148	30.67	0.9676	33.07	0.8993	28.89	0.8724	30.38	0.9029	30.21	0.9654	32.88	0.8891	28.55	0.8545	29.84
urban_0006	0.9131	31.82	0.9708	34.43	0.9130	31.32	0.8746	30.57	0.8973	31.17	0.9689	34.11	0.8972	30.67	0.8497	29.77
urban_0017	0.8777	31.46	0.9432	33.25	0.8952	30.94	0.8141	30.63	0.8606	30.98	0.9378	32.99	0.8819	30.56	0.7792	29.71
urban_0018	0.9301	30.70	0.9721	32.73	0.9274	28.53	0.8903	29.97	0.9174	30.20	0.9690	32.54	0.9188	28.21	0.8713	29.39
urban_0020	0.9062	32.58	0.9653	35.32	0.9242	30.85	0.8442	31.30	0.8895	31.94	0.9628	34.98	0.9146	30.43	0.8161	30.50
urban_0026	0.8786	28.66	0.9084	28.79	0.8492	27.63	0.8344	28.60	0.8661	28.36	0.9040	28.68	0.8387	27.40	0.8097	28.08
urban_0030	0.8754	30.29	0.9009	30.81	0.8401	28.31	0.8275	29.83	0.8629	29.85	0.8975	30.69	0.8267	28.08	0.8007	29.09
urban_0050	0.8687	28.01	0.9249	28.58	0.8722	27.09	0.7860	26.47	0.8514	27.79	0.9191	28.50	0.8535	26.85	0.7539	26.14
average	0.8956	30.52	**0.9441**	**32.12**	0.8901	29.20	0.8429	29.72	0.8810	30.07	**0.9406**	**31.92**	0.8776	28.85	0.8169	29.06

	$\sigma_{train} = 12$, $\sigma_{test} = 14.4$								$\sigma_{train} = 12$, $\sigma_{test} = 15.6$							
	Our basic		Our advanced		WASR [28]		DJF [27]		Our basic		Our advanced		WASR [28]		DJF [27]	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
urban_0004	0.8884	29.73	0.9625	32.57	0.8737	28.27	0.8233	29.03	0.8716	29.28	0.9595	32.39	0.8577	27.82	0.7981	28.38
urban_0006	0.8789	30.48	0.9662	33.84	0.8848	30.22	0.8264	29.05	0.8585	29.84	0.9627	33.52	0.8627	29.49	0.8022	28.33
urban_0017	0.8412	30.46	0.9330	32.77	0.8671	30.14	0.7474	29.07	0.8188	29.93	0.9261	32.49	0.8543	29.80	0.7156	28.40
urban_0018	0.9029	29.67	0.9652	32.33	0.9079	27.82	0.8479	28.70	0.8870	29.19	0.9620	32.09	0.8981	27.50	0.8306	28.16
urban_0020	0.8704	31.27	0.9562	34.61	0.9002	29.98	0.7824	29.66	0.8492	30.68	0.9524	34.37	0.8857	29.58	0.7608	29.04
urban_0026	0.8511	28.02	0.9022	28.61	0.8271	27.22	0.7877	27.67	0.8341	27.73	0.8985	28.52	0.8101	26.88	0.7606	27.19
urban_0030	0.8483	29.39	0.8953	30.52	0.8155	27.75	0.7785	28.44	0.8313	28.90	0.8903	30.27	0.8031	27.41	0.7498	27.76
urban_0050	0.8311	27.55	0.9152	28.42	0.8393	26.68	0.7229	25.83	0.8076	27.27	0.9116	28.37	0.8180	26.40	0.6915	25.53
average	0.8640	29.57	**0.9370**	**31.71**	0.8645	28.51	0.7896	28.43	0.8448	29.10	**0.9329**	**31.50**	0.8487	28.11	0.7636	27.85

dataset, as deep models can successfully take advantage of the availability of a huge amount of data for training. We use in-build parameter settings for DJF [27] and WASR [28] which have been optimum for training. Specifically, the training dataset for DJF [27] consists of 160,000 33×33 sub-images, for WASR [28] 50,000 sub-images of size 72×72 pixels, and for our advanced CDLSR 160,000 patches of size 8×8 pixels.[10]

As shown in Table IV, given a rich training dataset, deep models DJF [27], WASR [28] and our advanced CDLSR outperform shallow models such as JBF [23], GF [24], SDF [25], JFSM [26] and the basic version of our method as shown in Table I and Table II. Owing to the introduced multi-stage structure and neighbourhood regression techniques, the advanced version of our method can also take advantage of a large amount of data, yielding better performance in relation to the competing approaches. Moreover, we see a significant improvement over the basic version of our algorithm. For example, the average PSNR for 4 $\times$ super-resolution increases to 35.22 dB, leading to around 1.7 dB improvement over our basic version. Similarly, the average PSNR for 6 $\times$

[10] We need to highlight that the patch size does not have effect on the number of parameters of CNN based networks, such as the ones in DJF [27] and WASR [28], but the patch size may have impact on the number of parameters in other deep neural network architectures such as fully-connected neural networks. This issue is further explored in the supplementary materials.



(a) SSIM w.r.t. Testing Noise

(b) PSNR w.r.t. Testing Noise

Figure 8: Comparison with "deep" models on 4× upscaling of noisy LR testing near-infrared images with RGB modality for guidance. The training noise $\sigma_{train} = 0$ and the testing noise σ_{test} ranges from 2 to 12.

super-resolution increases to 33.41 dB with around 2.1 dB improvement over our basic version. The visual performance is shown in Figure 7 where it can be seen that our advanced version produces cleaner and sharper estimation than competing approaches. In addition, all these multimodal image SR deep models outperform single-modal image SR models, such as SRCNN [14], FSRCNN [15], CSC-Net [18].

2) Noisy Situation: More importantly, the other advantage of our approaches relates to the robustness in the presence of noise at training and/or testing stages, which is very common in practice [51], [52]. Here, we repeat the previous NIR-SR experiments as in Section V-B1 to test the robustness of competing algorithms in the presence of contamination of additive zero-mean Gaussian noise. The parameter setting is the same as in Section V-B1. Each evaluation metric value is averaged on all the testing images. We consider two typical scenarios:[11]

Noisy LR testing images. The first scenario assumes that the LR testing images are contaminated by zero-mean Gaussian noise with a certain standard deviation. Then, the models of WASR [28], DJF [27] and ours were all trained on the same noiseless training images and then tested on the same noisy LR images. In Table V and Figure 8, the results corresponding to setting $\sigma_{test} = \sigma_{train} = 0$ show that DJF [27] usually outperforms the basic version of our approach in the noise-free scenario, but does not surpass our advanced version. Moreover, other results corresponding to setting $\sigma_{test} \neq 0$ show that our algorithms demonstrate reasonable stability and robustness to noise, especially to strong noise. In contrast, DJF [27] is susceptible to noise and its performance degrades faster than ours. WASR [28] also demonstrates better robustness than DJF [27]. We believe that the good robustness and stability is due to sparsity priors exploited by the proposed model. In Figure 9, it is observed that the upscaling results of DJF [27] can not attenuate noise effectively, whereas our reconstruction is much cleaner.

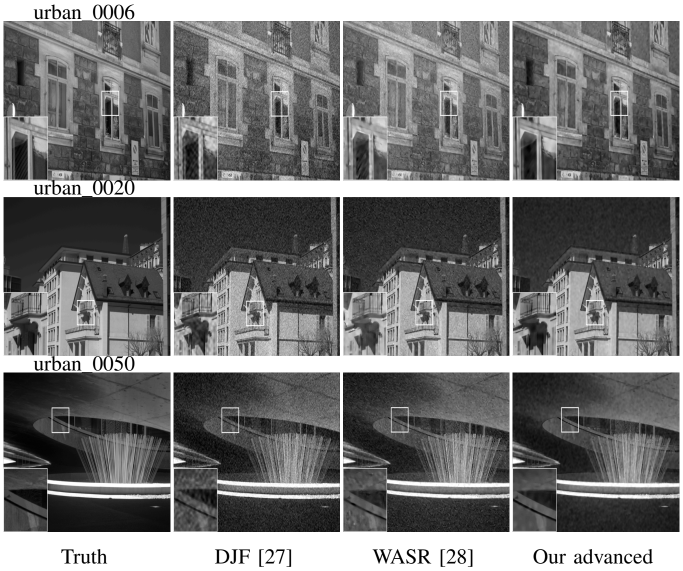


Figure 9: Comparison with "deep" models on 4× upscaling of noisy LR testing near-infrared images with RGB modality for guidance. The training noise $\sigma_{train} = 0$ and the testing noise $\sigma_{test} = 12$.

Noisy both LR testing and LR training images. The second scenario assumes that both the LR testing and the LR training images are contaminated by zero-mean Gaussian noise with a certain standard deviation. Then, the models of WASR [28], DJF [27] and ours were all trained on the same noisy training images and then tested on the same noisy LR images.

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 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× upscaling of near-infrared house images using and DJF [27], WASR [28], the basic and advanced version of our method. As shown in Table VI and Figure 11, the performance of all the algorithms degrades as the mismatch increases. However, the proposed algorithms not only has a slower degradation in performance than DJF [27] and WASR [28], but also yields higher SSIM and PSNR values. Moreover, our advanced version outperforms other methods with significant gains, illustrating our models are resilient to mismatched noise at training and testing phases.

3) Complexity: The proposed approach has other advantages with respect to DJF [27] and WASR [28]. One advantage relates to the amount of training time required by our approach. For example, DJF [27] takes about 12 hours to

[11] We 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.

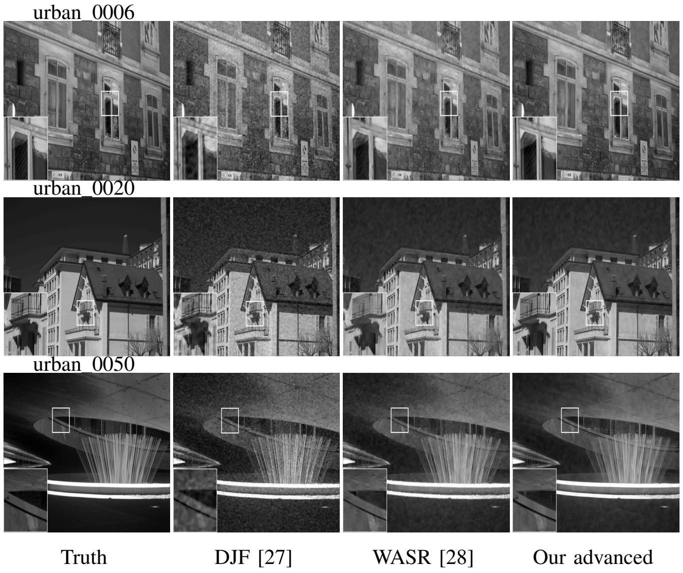
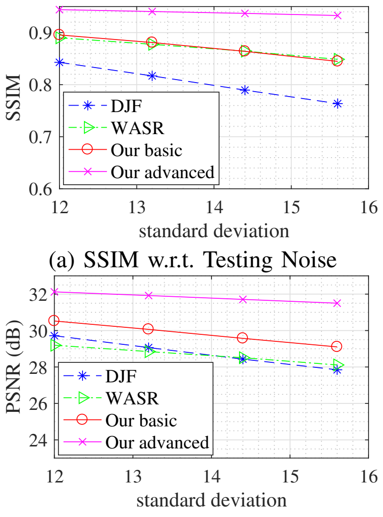


Figure 10: Comparison with "deep" models on 4× upscaling of noisy LR testing near-infrared images with RGB modality for guidance. The training noise $\sigma_{train} = 12$ and the testing noise $\sigma_{test} = 14.4$.



(a) SSIM w.r.t. Testing Noise

(b) PSNR w.r.t. Testing Noise

Figure 11: Comparison with "deep" models on 4× upscaling of noisy LR testing near-infrared images with RGB modality for guidance. The training noise $\sigma_{train} = 12$ and the testing noise $\sigma_{test} = [12, 13.2, 14.4, 15.6]$, corresponding to mismatch $\delta = [0, 10\%, 20\%, 30\%]$.

train through 50 epochs with an NVIDIA Titan black GPU for acceleration, and WASR [28] takes almost 11 hours to run through 10 stages with maximum 50 iterations in each stage. Contrary, the basic version of our approach takes only 118.2 seconds for training a group of coupled dictionaries on a computer equipped with a quadro-core i7 CPU at 3.4GHz with 32GB of memory, without any GPU acceleration. We note, however, that our basic version is slower than DJF and WASR during the testing phase. For example, it takes 117.8 seconds for our basic version to super-resolve an image of size 700 × 800 pixels, while only 13.6 seconds and 1.7 seconds for WASR and DJF, respectively. This is because we solve a non-convex optimization problem while they only perform a simple forward pass. But, the inference speed is improved by replacing time-consuming sparse coding with neighbourhood regression considerably, reducing the testing time to 30.7 seconds. We also need to mention that the acceleration of the inference speed is at the expense of increased training time due to the computation of the coupled sub-dictionaries for each neighbourhood. However, the training time of our advanced version is 3389 seconds (around 56 minutes), still far less than that of DJF [27] and WASR [28]. In addition, the acceleration of the inference speed owing to the neighborhood regression does not compromise the reconstruction performance. On the contrary, it improves the reconstruction performance. We also point out that our code was not optimized for speed. So, it might be possible to improve the inference speed further, for example, by extracting patches in parallel, optimizing some loop functions, rewriting or mex-compiling the code in C++ language.

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 complex dependency 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. In addition, by employing sparsity priors and taking both similarities and disparities into consideration, our approaches show decent robustness to structure inconsistency among different image modalities and possible (mismatched) noise in the LR testing and training images.

VI. 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 complex dependency relationship 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. By integrating multi-stage structure and neighbourhood regression techniques, our advanced version further improves the performance of the proposed model. Multispectral/RGB and NIR/RGB multimodal image SR experiments demonstrate that the proposed design brings notable benefits over state-of-the-art "shallow" models and "deep" models. Moreover, our methods also demonstrate remarkable robustness in the noisy situations where LR testing and/or training images are contaminated by mismatched noise.

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] 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.

[19] 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.

[20] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, 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.

[21] 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.

[22] 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.

[23] 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.

[24] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2010, pp. 1–14.

[25] B. Ham, M. Cho, and J. Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[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] 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.

[28] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, "Learning dynamic guidance for depth image enhancement," in *Proc. IEEE Conf. Comput. Vision Pattern Recog*, 2017, pp. 3769–3778.

[29] Q. Zhang, X. Shen, L. Xu, and J. Jia, "Rolling guidance filter," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2014, pp. 815–830.

[30] 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.

[31] 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.

[32] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[33] 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.

[34] 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.

[35] 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.

[36] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[37] 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.

[38] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Int. Conf. Comput. Vision*. IEEE, 1998, pp. 839–846.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.