COUPLED ISTA NETWORK FOR MULTI-MODAL IMAGE SUPER-RESOLUTION

Xin Deng and Pier Luigi Dragotti

School of Electrical and Electronic Engineering, Imperial College London.

ABSTRACT

In this paper, we propose a novel deep neural network architecture for multi-modal image super-resolution (MISR). The architecture is based on a new joint multi-modal dictionary learning (JMDL) algorithm to model cross-modality dependency and to map them to a high-resolution version of one modality. In JMDL, we learn three dictionaries and two transform matrices to combine the modalities. By using the learned model, we then design the network architecture by a coupled unfolding of the iterative shrinkage and thresholding algorithm (ISTA). We finally initialize the parameters of our network with a new optimization strategy. The initialized parameters are demonstrated to effectively decrease the training loss and increase the reconstruction accuracy. The numerical results show that our method outperforms other state-of-the-art methods quantitatively and qualitatively for MISR.

Index Terms— multi-modal image super-resolution, ISTA, dictionary learning, neural network.

1. INTRODUCTION

Single image super-resolution (SISR) is a typical problem in computer vision and image processing, which aims to infer a high-resolution (HR) image from a single low-resolution (LR) image. Many methods have been proposed to tackle this problem, including the methods based on dictionary learning [1, 2], and the more recent approaches based on deep networks [3, 4]. However, these methods only focus on the unimodal scenario, i.e., the LR and HR images are from the same modality.

Often, one scene is captured by multiple sensors, because information fused by different sensors can represent the scene more comprehensively. For example, in 3D model representation, both RGB and depth images are captured [5]. In remote sensing, multiple images are captured with different spectral bands. However, due to the limitations of storage capacity and the sensor mechanism, some images are captured with very low resolution, e.g., the depth images. Multi-modal image super-resolution (MISR) aims to improve the resolution of these images with the guidance of another HR image from a different modality. Some works [6, 7, 8] use multi-modal/joint dictionary learning to address this problem, but the requirement of computing the sparse codes makes these algorithms time-consuming. Other papers use deep neural networks to achieve the upscaling of one modality with the aid of another modality [9, 10, 11]. However, they use fully connected convolutional neural networks (CNN) which are not specifically designed for the multimodal scenario, and these deep networks might be difficult to interpret and train.

In this paper, we use a model-based approach to derive the architecture of a deep network for MISR. We first introduce a novel joint multi-modal dictionary learning (JMDL) algorithm to model the cross-modality dependencies. Then, based on the JMDL model, we design a new coupled deep network by unfolding the iterative shrinkage and thresholding algorithm (ISTA). Leveraging results in JMDL and the specific structure of the network, we devise an optimization strategy to initialize the parameters of the network before running the traditional back-propagation strategy. The end result is a simpler architecture easier to train but that outperforms state-of-the-art methods for MISR.

2. RELATED WORK

MISR. The MISR approaches can be broadly classified into two categories: joint image filtering based methods [12, 13, 14] and deep learning based methods [9, 10, 11]. The basic idea of joint image filtering is to transfer the salient structures in the guidance image, e.g., edges and textures, to the target image through constructing some joint filters. According to the filter type, joint image filtering methods can be further classified into two categories: static filtering [12] and dynamic filtering [14]. Recent works [9, 10, 11] proposed to use deep neural networks to solve this problem. Specifically, Li *et.al* [10] proposed to use CNN to achieve the upscaling of a LR image with a guided HR image from a different modality. The two works [9] and [11] proposed to super-resolve the depth image with the aid of the RGB image. However, these networks have the same disadvantages, i.e., their network architectures are designed empirically and what is happening inside is difficult to interpret. Moreover, their network parameters are all initialized randomly.

Iterative unfolding strategy. The iteration of many model-based parameter estimation algorithms usually consists of a linear operation followed by a non-linear thresholding, which is similar to the layer in a deep neural network. These iterative algorithms include, for example, ISTA [15] for sparse estimation, the approximate message passing (AMP) [16] for compressive sensing, the alternating direc-

Thanks to CSC Imperial Scholarship for funding.

tion method of multipliers (ADMM) [17] for generic inverse problem. Intuitively, we can turn the traditional iterative algorithms into interpretable deep networks by unfolding each iteration. Some works that use this unfolding perspective have appeared recently. For example, for the task of sparse codes estimation, the papers [18, 19] unroll the ISTA algorithm to be a deep network and [20] turns the AMP algorithm to be a deep network. Recently, Yang *et.al* [21] proposed to unfold the ADMM algorithm for compressive sensing magnetic resonance imaging (MRI), and Bertocchi *et.al* [22] proposed to unfold a proximal interior point method to a deep network for solving image deblurring problem.

3. PROPOSED METHOD

3.1. Joint multi-modal dictionary learning (JMDL)

The MISR task aims to find the HR patch $\boldsymbol{z}$ from the LR patch $\boldsymbol{x}$ with the guidance of HR patch $\boldsymbol{y}$, where $\boldsymbol{z}$ and $\boldsymbol{x}$ are from the same modality and $\boldsymbol{y}$ is from a different modality. We assume that $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are sparse in dictionaries $\boldsymbol{D}_x$, $\boldsymbol{D}_y$ and $\boldsymbol{D}_z$, respectively, and their sparse representations are correlated by two transform matrices $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$. Then, the JMDL problem can be formulated as follows:

$$\begin{aligned}\min_{\substack{\boldsymbol{D}_x,\boldsymbol{D}_y,\boldsymbol{D}_z,\\ \{\boldsymbol{\Lambda}_x,\boldsymbol{\Lambda}_y,\boldsymbol{\Lambda}_z,\}\\ \boldsymbol{W}_x,\boldsymbol{W}_y}} & \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}_x\boldsymbol{\Lambda}_x\|_F^2+\frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{D}_y\boldsymbol{\Lambda}_y\right\|_F^2+\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{D}_z\boldsymbol{\Lambda}_z\|_F^2\\ &+\lambda_x\|\boldsymbol{\Lambda}_x\|_1+\lambda_y\left\|\boldsymbol{\Lambda}_y\right\|_1+\lambda_z\|\boldsymbol{\Lambda}_z\|_1+\mu_x\|\boldsymbol{W}_x\|_F^2\\ &+\mu_y\left\|\boldsymbol{W}_y\right\|_F^2+\gamma\left\|\boldsymbol{\Lambda}_z-\boldsymbol{W}_x\boldsymbol{\Lambda}_x-\boldsymbol{W}_y\boldsymbol{\Lambda}_y\right\|_F^2,\\ & s.t.,\left\|\boldsymbol{d}_{x,i}\right\|_2^2\le 1,\left\|\boldsymbol{d}_{y,i}\right\|_2^2\le 1,\left\|\boldsymbol{d}_{z,i}\right\|_2^2\le 1,\forall i\end{aligned} \tag{1}$$

where $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}\in\mathbb{R}^{n\times t}$ are the matrices related to the gathered training samples, and λ_x, λ_y and λ_z are the regularization parameters for the sparse representations $\boldsymbol{\Lambda}_x$, $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_z$, respectively. Moreover, μ_x and μ_y are the regularization parameters for the transform matrices, and γ is the regularization parameter for the sparse representation mapping error. Finally, $\boldsymbol{d}_{x,i}$, $\boldsymbol{d}_{y,i}$ and $\boldsymbol{d}_{z,i}$ are the i-th atom of dictionaries $\boldsymbol{D}_x$, $\boldsymbol{D}_y$, and $\boldsymbol{D}_z$, respectively. The two most related models with Eq. (1) are the SCDL model in [23] and the $SliM^2$ model in [24]. However, the SCDL model only correlate two modalities and the $SliM^2$ model only establishes the single mapping from one modality to another. In contrast, our JMDL model establish the joint mapping from two modalities to a third modality.

This problem in Eq. (1) is not convex with regard to $\boldsymbol{D}_x$, $\boldsymbol{D}_y$, $\boldsymbol{D}_z$, $\boldsymbol{\Lambda}_x$, $\boldsymbol{\Lambda}_y$, $\boldsymbol{\Lambda}_z$, $\boldsymbol{W}_x$ and $\boldsymbol{W}_z$. However, it is convex to one variable when the others are fixed. Thus, we can solve this problem using an alternating method. We first fix the dictionaries and transform matrices to update the sparse representations, then we fix the sparse representations and the transform matrices to update the dictionaries, and finally we fix the dictionaries and sparse representations to learn the transform matrices.

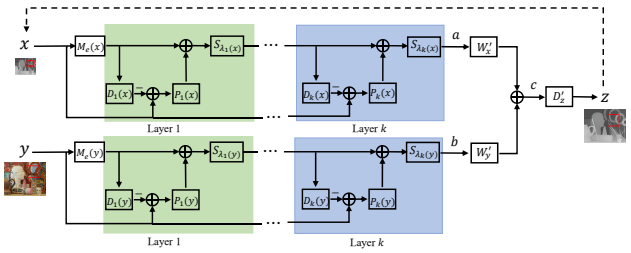


Fig. 1: The architecture of the proposed coupled ISTA network.

3.2. Coupled ISTA network

In the synthesis phase, given a LR patch $\boldsymbol{x}$ and a guided HR patch $\boldsymbol{y}$, we first need to find the sparse coefficients with the learned dictionaries $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ through solving:

$$\min_{\{\boldsymbol{a},\boldsymbol{b}\}}\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{D}_x\boldsymbol{a}\|_2^2+\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{D}_y\boldsymbol{b}\right\|_2^2+\lambda_x\|\boldsymbol{a}\|_1+\lambda_y\|\boldsymbol{b}\|_1. \tag{2}$$

In Eq. (2), the updating of $\boldsymbol{a}$ and $\boldsymbol{b}$ is independent of each other, which are both LASSO problems. After obtaining $\boldsymbol{a}$ and $\boldsymbol{b}$, we can have sparse representation $\boldsymbol{c}=\boldsymbol{W}_x\boldsymbol{a}+\boldsymbol{W}_y\boldsymbol{b}$. Finally, the HR patch $\boldsymbol{z}$ can be calculated by multiplying $\boldsymbol{c}$ by the dictionary $\boldsymbol{D}_z$.

The aforementioned algorithm has two drawbacks. Firstly, the calculation of sparse representations $\boldsymbol{a}$ and $\boldsymbol{b}$ relies on an iterative algorithm, e.g., ISTA or FISTA, which is time-consuming. Secondly, the synthesis and training phases are not fully correlated with each other, i.e., the ground-truth HR patch is not accessible in the synthesis phase, which may decrease the reconstruction accuracy. To overcome these drawbacks, we propose a coupled ISTA network by unfolding the ISTA algorithm. The architecture of the network is shown in Fig. 1. Specifically, the network is composed of two branches: the upper branch aims to infer the sparse representation $\boldsymbol{a}$ for the LR input patch, while the lower branch aims to infer the sparse representation $\boldsymbol{b}$ for the guided HR patch. Take the upper branch for example, the ISTA algorithms works in iterations to obtain $\boldsymbol{a}$ as follows:

$$\boldsymbol{a}_k=S_{\lambda_k}(\boldsymbol{a}_{k-1}+\boldsymbol{D}_x^T(\boldsymbol{x}-\boldsymbol{D}_x\boldsymbol{a}_{k-1})), \tag{3}$$

where $\boldsymbol{a}_k$ is the value of $\boldsymbol{a}$ at the k-th iteration. Through unfolding the Eq. (3), we can have the upper branch in Fig. 1. In order to make the network more flexible, we make three relaxations about the original ISTA algorithm. Firstly, the dictionary $\boldsymbol{D}_x$ is not required to be the same across different layers, i.e., we have a set of synthesis dictionaries $\{\boldsymbol{D}_1(x),\cdots,\boldsymbol{D}_k(x)\}$. Secondly, the relationship between $\boldsymbol{D}_x^T$ and $\boldsymbol{D}_x$ is broken, instead we have another set of analysis dictionaries $\{\boldsymbol{P}_1(x),\cdots,\boldsymbol{P}_k(x)\}$ as shown in Fig. 1. Thirdly, the soft threshold is allowed to change across layers, and we use a vector threshold instead of a constant scalar. We have different vector thresholds $\{\boldsymbol{\lambda}_1(x),\cdots,\boldsymbol{\lambda}_k(x)\}$ for each layer.

The lower branch can be obtained in the same way. Then, the outputs of these two branches are combined by the transform matrices to obtain $\boldsymbol{c}$ which is further multiplied by the reconstruction matrix to reconstruct the HR patch $\boldsymbol{z}$.

3.3. Layer-wise initialization algorithm.

Before training the deep network, we propose a layer-wise initialization algorithm to initialize all the network parameters. Take the upper branch for example, we aim to minimize the mean squared error (MSE) between the predicted sparse codes by the upper branch and the target sparse codes $\boldsymbol{\Lambda}_x$ obtained by solving (1). Since the parameters of the upper and lower branches can be initialized using the same algorithm, we just ignore the subscript x and y to make the notations simpler. Specifically, in the k-th layer, we have the following optimization target:

$$\{\boldsymbol{P}_k, \boldsymbol{D}_k, \boldsymbol{\lambda}_k\} = \underset{\boldsymbol{P}_k, \boldsymbol{D}_k, \boldsymbol{\lambda}_k}{\operatorname{argmin}} \left\| \boldsymbol{\Lambda} - S_{\boldsymbol{\lambda}_k}(\boldsymbol{B}_{k-1} + \boldsymbol{P}_k(\boldsymbol{X} - \boldsymbol{D}_k\boldsymbol{B}_{k-1})) \right\|_F^2, \tag{4}$$

where $\boldsymbol{B}_{k-1}$ is the predicted sparse code of the previous $(k-1)$-th layer. Here, $\boldsymbol{P}_k \in \mathbb{R}^{m\times n}(m > n)$ can be regarded as an analysis dictionary and $\boldsymbol{D}_k \in \mathbb{R}^{n\times m}$ is a synthesis dictionary. $\boldsymbol{\lambda}_k \in \mathbb{R}^m$ is a threshold vector of the k-th layer. We initialize $\boldsymbol{B}_0$ by minimizing the reconstruction error of training samples $\boldsymbol{X}$ using the original dictionary $\boldsymbol{D}$ (equal to $\boldsymbol{D}_x$ for the upper branch and $\boldsymbol{D}_y$ for the lower branch):

$$\boldsymbol{B}_0 = \underset{\boldsymbol{B}_0}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{B}_0\|_F^2 + \mu \|\boldsymbol{B}_0\|_F^2, \tag{5}$$

which can be solved by ridge regression to get solution as $\boldsymbol{B}_0 = (\boldsymbol{D}^T\boldsymbol{D} + \mu\boldsymbol{I})^{-1}\boldsymbol{D}^T\boldsymbol{X}$, where $\boldsymbol{I}$ is the identity matrix. To make the network structure simple, we define a mapping matrix $\boldsymbol{M}_e \in \mathbb{R}^{m\times n}$ which directly maps $\boldsymbol{X}$ to be $\boldsymbol{B}_0$, and we initialize $\boldsymbol{M}_e$ as $(\boldsymbol{D}^T\boldsymbol{D} + \mu\boldsymbol{I})^{-1}\boldsymbol{D}^T$. Next, we focus on solving the optimization problem in (4). Since we have three variables to be optimized, we use an alternating way to update them iteratively.

Step 1. We fix $\boldsymbol{D}_k$ and $\boldsymbol{\lambda}_k$ to update the analysis dictionary $\boldsymbol{P}_k$. In this case, since $(\boldsymbol{X} - \boldsymbol{D}_k\boldsymbol{B}_{k-1})$ does not change, we denote it by $\boldsymbol{U}$, and then $\boldsymbol{P}_k$ can be obtained through solving the following optimization:

$$\boldsymbol{P}_k = \underset{\boldsymbol{P}_k}{\operatorname{argmin}} \left\| \boldsymbol{\Lambda} - S_{\boldsymbol{\lambda}_k}(\boldsymbol{B}_{k-1} + \boldsymbol{P}_k\boldsymbol{U}) \right\|_F^2. \tag{6}$$

We update the atoms of $\boldsymbol{P}_k$ row by row. When updating the j-th atom, the other atoms remain fixed. Specifically, the j-th atom of $\boldsymbol{P}_k$ is updated by

$$\boldsymbol{p}_j^T = \underset{\boldsymbol{p}_j^T}{\operatorname{argmin}} \left\| \boldsymbol{g}_j^T - S_{\lambda_k^j}(\boldsymbol{b}_j^T + \boldsymbol{p}_j^T\boldsymbol{U}) \right\|_2^2, \tag{7}$$

where $\boldsymbol{g}_j^T$, $\boldsymbol{b}_j^T$, $\boldsymbol{p}_j^T$ are the j-th $(0 < j \le m)$ row of $\boldsymbol{\Lambda}$, $\boldsymbol{B}_{k-1}$, and $\boldsymbol{P}_k$, respectively. λ_k^j is the j-th element in vector $\boldsymbol{\lambda}_k$. The difficulty here is that the soft-thresholding operator is non-linear. Inspired by [25], we divide the non-linear operation into two linear operations. Actually, the soft-thresholding operator splits the signals $(\boldsymbol{b}_j^T + \boldsymbol{p}_j^T\boldsymbol{U})$ into two sets: one set with all zeros after soft-thresholding and the other set with non-zero values. Suppose $\boldsymbol{J}$ denotes the indices of the non-zero samples and $\hat{\boldsymbol{J}}$ denotes the indices of samples that are set to zero, we can split $\boldsymbol{U}$ into $\boldsymbol{U}^J$ and $\boldsymbol{U}^{\hat{J}}$. Likewise, we can split $\boldsymbol{g}_j^T$ into $\boldsymbol{g}_j^J$ and $\boldsymbol{g}_j^{\hat{J}}$, $\boldsymbol{b}_j^T$ into $\boldsymbol{b}_j^J$ and $\boldsymbol{b}_j^{\hat{J}}$. Then, (7) can be written as follows,

$$\boldsymbol{p}_j^T = \underset{\boldsymbol{p}_j^T}{\operatorname{argmin}} \left\| \boldsymbol{g}_j^J - (\boldsymbol{b}_j^J + \boldsymbol{p}_j^T\boldsymbol{U}^J \pm \lambda_k^j) \right\|_2^2 + \left\| \boldsymbol{g}_j^{\hat{J}} \right\|_2^2. \tag{8}$$

Here, we assume that $\left\| \boldsymbol{g}_j^{\hat{J}} \right\|_2^2$ is constant when the threshold is fixed, and we can simplify (8) to be

$$\boldsymbol{p}_j^T = \underset{\boldsymbol{p}_j^T}{\operatorname{argmin}} \left\| \boldsymbol{g}_j^J - \boldsymbol{b}_j^J - \boldsymbol{p}_j^T\boldsymbol{U}^J \pm \lambda_k^j \right\|_2^2, \tag{9}$$

which can be solved through least square fitting, and we can calculate $\boldsymbol{p}_j^T$ by

$$\boldsymbol{p}_j^T = (\boldsymbol{g}_j^J - \boldsymbol{b}_j^J \pm \lambda_k^j)(\boldsymbol{U}^J)^T(\boldsymbol{U}^J(\boldsymbol{U}^J)^T + \mu\boldsymbol{I})^{-1}, \tag{10}$$

where μ is a regularization parameter applied when $\boldsymbol{U}^J$ is not full rank.

Step 2. We fix $\boldsymbol{P}_k$ and $\boldsymbol{\lambda}_k$ to update $\boldsymbol{D}_k$, through solving the following optimization:

$$\boldsymbol{D}_k = \underset{\boldsymbol{D}_k}{\operatorname{argmin}} \left\| \boldsymbol{\Lambda} - S_{\boldsymbol{\lambda}_k}(\boldsymbol{B}_{k-1} + \boldsymbol{P}_k\boldsymbol{X} - \boldsymbol{P}_k\boldsymbol{D}_k\boldsymbol{B}_{k-1})) \right\|_F^2. \tag{11}$$

To solve this problem, we first denote $\boldsymbol{V}_k = \boldsymbol{P}_k\boldsymbol{D}_k$ and update $\boldsymbol{V}_k$ using the same procedure as in Step 1. Then, since we already have the updated version of $\boldsymbol{P}_k$, we can further spilt $\boldsymbol{D}_k$ from $\boldsymbol{V}_k$ and this yields

$$\boldsymbol{D}_k = (\boldsymbol{P}_k^T\boldsymbol{P}_k + \mu\boldsymbol{I})^{-1}\boldsymbol{P}_k^T\boldsymbol{V}_k. \tag{12}$$

Step 3. We fix $\boldsymbol{P}_k$ and $\boldsymbol{D}_k$ to update the thresholds $\boldsymbol{\lambda}_k$. Here, $\boldsymbol{\lambda}_k$ is a vector, and we update each element of it independently. The updating of the j-th element is through the following optimization:

$$\lambda_k^j = \underset{\lambda_k^j}{\operatorname{argmin}} \left\| \boldsymbol{g}_j^T - S_{\lambda_k^j}(\boldsymbol{b}_j^T + \boldsymbol{p}_j^T\boldsymbol{U}) \right\|_2^2. \tag{13}$$

Suppose that $\boldsymbol{q}^T = \boldsymbol{b}_j^T + \boldsymbol{p}_j^T\boldsymbol{U}$, and the elements in $\boldsymbol{q}^T$ are sorted such that $|q_1| \le |q_2| \le \ldots \le |q_t|$. The candidates for λ_k^j can be selected from the list $\{|q_1|/2, (|q_1|+|q_2|)/2, (|q_2|+|q_3|)/2, \ldots, (|q_t|+1)/2\}$. The value which minimizes the loss in (13) is the new updated value of λ_k^j.

These three steps are iterated until reaching the maximum iteration. After we obtain $\boldsymbol{P}_k$, $\boldsymbol{D}_k$, and $\boldsymbol{\lambda}_k$, we can update the new sparse codes $\boldsymbol{B}_k$ through:

$$\boldsymbol{B}_k = S_{\boldsymbol{\lambda}_k}(\boldsymbol{B}_{k-1} + \boldsymbol{P}_k(\boldsymbol{X} - \boldsymbol{D}_k\boldsymbol{B}_{k-1})). \tag{14}$$

Then, the parameter initialization of the next layer can be launched. Finally, after K layer initialization, we obtain the final predicted sparse codes $\boldsymbol{B}_K(x)$ for the upper branch and $\boldsymbol{B}_K(y)$ for the lower branch. Then, the two transform matrices can be initialized by:

$$\boldsymbol{W}_x' = \boldsymbol{W}_x\boldsymbol{\Lambda}_x\boldsymbol{B}_K^T(x)(\boldsymbol{B}_K(x)\boldsymbol{B}_K^T(x) + \mu\boldsymbol{I})^{-1}, \tag{15}$$

$$\boldsymbol{W}_y' = \boldsymbol{W}_y\boldsymbol{\Lambda}_y\boldsymbol{B}_K^T(y)(\boldsymbol{B}_K(y)\boldsymbol{B}_K^T(y) + \mu\boldsymbol{I})^{-1}, \tag{16}$$

Table 1: Effectiveness of the JMDL algorithm

Methods	Bicubic	CDL	JMDL
RMSE	6.10	4.95	4.86
SSIM	0.9536	0.9688	0.9702

where $\boldsymbol{W}_x$, $\boldsymbol{W}_y$, $\boldsymbol{\Lambda}_x$, and $\boldsymbol{\Lambda}_y$ are obtained by solving (1). By denoting $\boldsymbol{B}_K(z) = \boldsymbol{W}'_x\boldsymbol{B}_K(x) + \boldsymbol{W}'_y\boldsymbol{B}_K(y)$, we can initialize the reconstruction matrix $\boldsymbol{D}'_z$ as:

$$\boldsymbol{D}'_z = \boldsymbol{Z}\boldsymbol{B}_K^T(z)(\boldsymbol{B}_K(z)\boldsymbol{B}_K^T(z) + \mu\boldsymbol{I})^{-1}. \quad (17)$$

After initialization, we train the network in Fig. 1 using the standard back-propagation algorithm.

4. NUMERICAL RESULTS

The experiments are performed in the RGB and depth multi-modal scenario[1], including the Middlebury[2] and Sintel[3] datasets, for $4\times$ upscaling. For training, we extract 2,000,000 patches with size 8×8 from the training dataset provided by [26]. The number of layers K is 2 and the dictionary size m is 256. The Adam optimizer is used to train the network for 200 epochs with basic learning rate as 0.0001.

4.1. Effectiveness of JMDL algorithm

Table 1 compares the root mean squared error (RMSE) and structural similarity (SSIM) results of the JMDL algorithm with a recently proposed coupled dictionary learning (CDL) algorithm [8] for MISR task. The results are averaged among all testing images. We can see that our algorithm outperforms CDL in both RMSE and SSIM, validating the effectiveness of the new multi-modal model.

4.2. The role of initialization

We compare in Fig. 2 (a) the training loss curves with random initialization and our initialization algorithms. It can be seen that our algorithm yields a lower training loss. Moreover, we have a much lower loss start than the random one, since the parameters have already been optimized by our initialization algorithm. We also compare the RMSE performance in Fig. 2 (b) with these two initializations, and we can see that our algorithm constantly outperforms the random initialization in the reconstruction accuracy, independently of the number of epochs.

4.3. Comparison against other methods

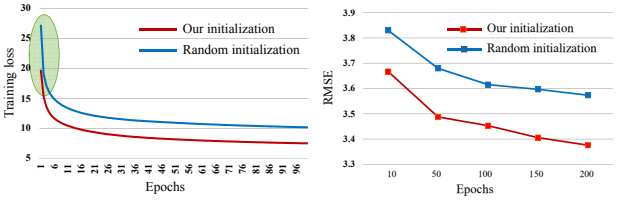
We compare our method with the following approaches for depth image super-resolution: Ferstl *et al.* [27], Xie *et al.* [28], Park *et al.* [29], Lu *et al.* [30], Gu *et al.* [31], Wang *et al.* [3], Kim *et al.* [4], and Song *et al.* [32]. The numerical results are compared by two measurements: RMSE and SSIM. Table 2 presents the RMSE and SSIM results for the Middlebury and Sintel datasets. In this table, our results are obtained by cascading the network in Fig. 1. We can see that our method outperforms the other state-of-the-art approaches.

Fig. 3 visualizes the $4\times$ upscaling results of image *Art* from the Middlebury dataset with different methods. As can be seen from this figure, our method reconstructs clearer and sharper edges than the other methods.

[1]Our method can provide state-of-the-art results in other multi-modal cases, but we only provide the results in RGB/depth case due to space limit.

[2]http://vision.middlebury.edu/stereo/data/scenes2005/

[3]http://sintel.is.tue.mpg.de/



(a) (b)

Fig. 2: (a) shows the training loss across 100 training epochs with random initialization and our initialization methods, and (b) compares the average RMSE value of testing images with these two initialization methods.

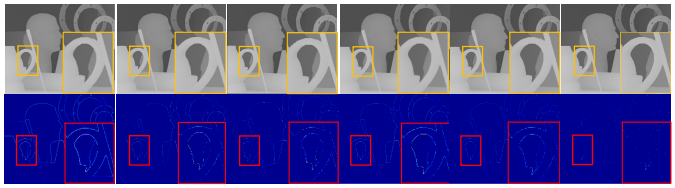


Fig. 3: Visual comparison of *Art* in Middlebury dataset with upscaling factor $= 4$. The first row shows the reconstructed depth images, and the second row shows the error maps. (a) Bicubic. (b) Ferstl *et al.* [27]. (c) Xie *et al.* [28]. (d) Gu *et al.*[31]. (e)Wang *et al.*[3]. (f) Our method.

5. CONCLUSION

In this paper, we introduced a joint multi-modal dictionary learning model for multi-modal image super-resolution task. Based on this model, we further proposed a new deep neural network through unfolding the ISTA algorithm. We also introduced a novel way to initialize all the network parameters by solving a multi-layer dictionary learning problem. Compared with the random initialization, our new initialization algorithm is demonstrated to achieve better performance in both training and testing phases. Numerical results show that our method improves significantly over other state-of-the-art methods in RGB/depth image super-resolution.

Table 2: Results on the Middlebury and Sintel datasets for $4\times$ upscaling, with the best results in bold.

Methods	Ambush		Bamboo		Cave		Market		Art		Books		Moebius	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
Bicubic	6.39	0.9685	14.09	0.8760	6.61	0.9503	8.83	0.9295	3.87	0.9687	1.60	0.9911	1.32	0.9908
Xie *et al.* [28]	8.79	0.9438	19.02	0.8301	9.14	0.9221	12.21	0.8869	3.79	0.9758	1.63	0.9917	1.33	0.9910
Park *et al.* [29]	6.03	0.9678	12.05	0.8910	7.13	0.9379	9.45	0.9067	3.76	0.9752	1.66	0.9912	1.42	0.9911
Ferstl *et al.* [27]	5.99	0.9701	11.54	0.8950	6.40	0.9563	8.01	0.9298	3.73	0.9771	1.65	0.9915	1.43	0.9909
Lu *et al.* [30]	5.53	0.9712	10.61	0.9028	6.10	0.9610	8.31	0.9266	4.10	0.9747	2.18	0.9896	1.56	0.9896
Gu *et al.*[31]	6.04	0.9766	13.35	0.9001	6.15	0.9613	8.10	0.9470	3.52	0.9779	1.57	0.9923	1.23	0.9930
Wang *et al.*[3]	4.29	0.9850	9.63	0.9389	4.37	0.9769	5.94	0.9664	2.59	0.9858	1.08	0.9951	0.93	0.9949
Kim *et al.*[4]	3.18	0.9913	9.18	0.9465	3.55	0.9839	5.52	0.9703	1.87	0.9926	0.75	0.9968	0.87	0.9952
Song *et al.*[32]	-	-	-	-	-	-	-	-	1.89	0.9889	0.92	0.9930	-	-
Our	**2.80**	**0.9928**	**8.13**	**0.9547**	**3.13**	**0.9861**	**4.85**	**0.9755**	**1.73**	**0.9936**	**0.70**	**0.9969**	**0.77**	**0.9959**

6. REFERENCES

[1] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

[2] Radu Timofte, Vincent De Smet, and Luc Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[3] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang, "Deep networks for image super-resolution with sparse prior," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 370–378.

[4] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[5] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Aligning 3d models to rgb-d images of cluttered scenes," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 4731–4740.

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

[7] Pingfan Song, Joao F.C Mota, Nikos Deligiannis, and Miguel R. D Rodrigues, "Coupled dictionary learning for multimodal image super-resolution," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2016.

[8] Pingfan Song, Xin Deng, João FC Mota, Nikos Deligiannis, Pier Luigi Dragotti, and Miguel RD Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *arXiv preprint arXiv:1709.08680*, 2017.

[9] Xibin Song, Yuchao Dai, and Xueying Qin, "Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 360–376.

[10] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep joint image filtering," in *European Conference on Computer Vision*. Springer, 2016, pp. 154–169.

[11] Gernot Riegler, David Ferstl, Matthias Rüther, and Horst Bischof, "A deep primal-dual network for guided depth super-resolution," *arXiv preprint arXiv:1607.08569*, 2016.

[12] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics (ToG)*, vol. 26, no. 3, pp. 96, 2007.

[13] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi, "Joint geodesic upsampling of depth images," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 169–176.

[14] Bumsub Ham, Minsu Cho, and Jean Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 192–207, 2018.

[15] Ingrid Daubechies, Michel Defrise, and Christine De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[16] David L Donoho, Arian Maleki, and Andrea Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[17] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[18] Karol Gregor and Yann LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omnipress, 2010, pp. 399–406.

[19] Zhangyang Wang, Qing Ling, and Thomas Huang, "Learning deep l0 encoders," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 2194–2200.

[20] Mark Borgerding, Philip Schniter, and Sundeep Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4293–4308, 2017.

[21] Jian Sun, Huibin Li, Zongben Xu, et al., "Deep ADMM-Net for compressive sensing mri," in *Advances in Neural Information Processing Systems*, 2016, pp. 10–18.

[22] Carla Bertocchi, Emilie Chouzenoux, Marie-Caroline Corbineau, Jean-Christophe Pesquet, and Marco Prato, *Deep Unfolding of a Proximal Interior Point Method for Image Restoration*, Ph.D. thesis, CVN, CentraleSupélec, Université Paris-Saclay, Gif-Sur-Yvette, France , 2018.

[23] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2216–2223.

[24] Yueting Zhuang, Yanfei Wang, Fei Wu, Yin Zhang, and Weiming Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval.," in *AAAI*, 2013, pp. 1070–1076.

[25] Ron Rubinstein and Michael Elad, "Dictionary learning for analysis-synthesis thresholding," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5962–5972, 2014.

[26] Gernot Riegler, Matthias Rüther, and Horst Bischof, "ATGV-net: Accurate depth super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 268–284.

[27] David Ferstl, Matthias Ruther, and Horst Bischof, "Variational depth superresolution using example-based edge representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 513–521.

[28] Jun Xie, Rogerio Schmidt Feris, and Ming-Ting Sun, "Edge-guided single depth image super resolution," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 428–438, 2016.

[29] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1623–1630.

[30] Jiajun Lu and David Forsyth, "Sparse depth super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2245–2253.

[31] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang, "Learning dynamic guidance for depth image enhancement," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 10, no. y2, pp. 2, 2017.

[32] Xibin Song, Yuchao Dai, and Xueying Qin, "Deeply supervised depth map super-resolution as novel view synthesis," in *IEEE Transactions on circuits and systems for video technology*, 2018, p. Early access.