Multi-modal Image Super-resolution with Joint Coupled Deep Transform Learning

R Krishna Kanth*, Andrew Gigie*, Kriti Kumar†*, A Anil Kumar*, Angshul Majumdar†, Balamuralidhar P*
*TCS Research, India, † IIIT Delhi, New Delhi, India.
Email: {rokkamkrishna.kanth, gigie.andrew, kriti.kumar, achannaanil.kumar, balamurali.p}@tcs.com,
angshul@iiitd.ac.in

***Abstract*—In this paper, we address the problem of multi-modal image super-resolution (MISR), which aims at improving the resolution of the target modality with the help of high resolution guidance image of another modality. A novel joint coupled deep transform learning framework (JCDTL) based on deep transform learning is proposed which combines the information from multiple modalities for achieving MISR. The formulation and the requisite solution steps are provided. Two publicly available datasets RGB/NIR and RGB/Multispectral are considered for performance evaluation. The proposed approach shows a considerable improvement in performance compared to the state-of-art techniques. Further, an average PSNR improvement of close to 2dB and 1.5dB on RGB/Multispectral and RGB/NIR datasets respectively is observed by increasing the number of layers from one to three.**

***Index Terms*—Multi-modal image super-resolution, Transform Learning, joint optimization, Deep Transform Learning.**

I. Introduction

Multi-modal imaging systems provide enriched information and hence are often employed in several applications such as remote sensing [1] [2], seed viability studies [3], environment monitoring [4], food processing [5], medical field [6] and forensic studies [7]. While they provide several benefits, these multi-modal systems will have different resolutions due to design complexity and/or limitations posed by physical constraints. For instance, in a RGB–hyperspectral imaging system that is commonly used in remote sensing, the RGB will have a better spatial resolution and poorer spectral resolution, and vice-versa for the hyperspectral imaging system. The necessity to thus enhance the resolution is not only essential but may also be required in several downstream applications. Unlike the uni-modal imaging system, the multi-modal imaging system can exploit the cross-modal information between different modalities to enhance the resolution in a better way.

Multi-modal Image Super-Resolution (MISR) aims to enhance resolution by leveraging cross-modal information. The idea here is to enhance the resolution of a Low Resolution (LR) imaging modality (*target modality*) with the help of another Higher Resolution (HR) modality (*guidance modality*) as they share some common features like edges, textures, etc. The interest in MISR has grown recently, and the techniques available in literature can broadly be classified into: i) Classical signal/image processing based techniques and ii) Data-driven techniques. The papers [8]-[10], belong to the first category, where joint image-based filtering techniques are employed by constructing joint filters taking into account certain features like edges and textures from the guidance modality. However, as shown in [11] these techniques fail when disparity between the target and guidance modality exists. On the other hand, in the data-driven techniques, the cross-modal information is learned from training samples. The popular Deep Learning (DL) based approaches employing CNNs have been used for MISR in [12]-[14]. However, they require huge amount of training data for better reconstruction, and may not perform well for limited training data scenario, as also observed in[11].

The work in [11] proposed a dictionary learning based approach where separate and common dictionaries are learnt for different modalities with the assumption that the common dictionary shares the same sparse representation. In [15], by modeling the cross-modal dependencies as a weighted superposition of individual sparse dictionary coefficients, they presented a Joint Multi-modal Dictionary Learning (JMDL) approach for MISR. Transform Learning (TL) [16] which is the analysis counterpart of dictionary learning, is gaining a lot of prominence these days due to their computational advantage and lesser sparsification error. Driven by the advantages of TL, very recently in [17], along similar lines as JMDL, Joint Coupled Transform Learning (JCTL) was presented by the authors of this paper, which showed better reconstruction performance compared to other techniques.

In this paper, motivated by the advantages of TL and performance of JCTL [17] for MISR, we propose a deeper version of JCTL referred to as Joint Coupled Deep Transform Learning (JCDTL). Unlike JCTL, where only 1-layer of Transform is employed for both the guidance and target modality, in JCDTL, deeper layers of TL, i.e., Deep Transform Learning (DTL) [18] is employed at both the guidance and target modality. The TL coefficients of the target HR modality are modeled as the weighted superposition of the sparse DTL coefficients of the target LR modality and the guidance HR modality. A novel JCDTL optimization framework is provided, relating the various deep transforms, their corresponding coefficients, and the superposition weights.

The results obtained with two publicly available datasets, namely, RGB/NIR [19] and RGB/Multispectral [20], demonstrates the improved performance of the proposed JCDTL method compared to the stat-of-the-art methods. Further, by increasing the number of layers from 1 to 3, a PSNR improvement of around 1.5dB and 2dB is observed for RGB/NIR and RGB/Multispectral data, respectively.

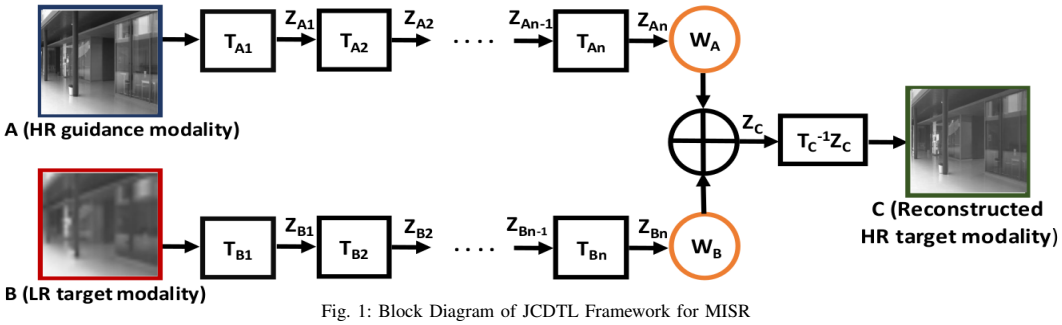

Fig. 1: Block Diagram of JCDTL Framework for MISR

The rest of the paper is organized as follows. Section II presents a brief background on TL and DTL for the sake of completeness. Section III presents the proposed JCDTL formulation with the requisite solution steps. This is followed by Section IV that presents the numerical results and finally, Section V concludes the paper.

II. Background on Transform Learning

This section presents a brief background on Transform Learning (TL) [21] and its deep version, DTL [18] that will be employed in our proposed MISR formulation. Given an input $\mathbf{A} \in \mathbb{R}^{F \times M}$ where F is the number of features and M the number of measurements, the TL is expressed as: $\mathbf{TA} = \mathbf{Z}$ where, $\mathbf{T} \in \mathbb{R}^{K \times F}$ is the transform with K atoms and $\mathbf{Z} \in \mathbb{R}^{K \times M}$ are the associated coefficients [16]. The modified learning formulation that enforces sparsity on $\mathbf{Z}$ is given as [22]:

$$\min_{\mathbf{T},\mathbf{Z}} \|\mathbf{TA} - \mathbf{Z}\|_F^2 + \lambda(\|\mathbf{T}\|_F^2 - \log\det \mathbf{T}) + \gamma(\|\mathbf{Z}\|_0) \quad (1)$$

where the constraint $(\|\mathbf{T}\|_F^2 - \log\det \mathbf{T})$ is added to prevent trivial solution and also to control the condition number of the learnt transform $\mathbf{T}$. The optimization problem in (1) is non convex with respect to $\mathbf{T}$ and $\mathbf{Z}$ jointly, hence, variable splitting and ADMM technique is employed to solve for $\mathbf{T}$ and $\mathbf{Z}$ [23]. $\mathbf{T}$ is updated by solving the following:

$$\min_{\mathbf{T}} \|\mathbf{TA} - \mathbf{Z}\|_F^2 + \lambda(\|\mathbf{T}\|_F^2 - \log\det \mathbf{T}) \quad (2)$$

The closed-form update for $\mathbf{T}$ is obtained using Cholesky decomposition given as: $\mathbf{AA}^T + \lambda\mathbf{I} = \mathbf{LL}^T$, where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{L}^T$ denotes the conjugate transpose of $\mathbf{L}$. Subsequently, SVD on $\mathbf{L}^\dagger\mathbf{AZ}^T = \mathbf{QSR}^T$, where, $\dagger$ denotes pseudo-inverse, the diagonal entries of $\mathbf{S}$ are the singular values and $\mathbf{Q}$, $\mathbf{R}$ are the left and right singular vectors of $(\mathbf{L}^\dagger\,\mathbf{AZ}^T)$ respectively. This results in the following update for $\mathbf{T}$ [24]:

$$\mathbf{T} = 0.5\mathbf{R}(\mathbf{S} + (\mathbf{S}^2 + 2\lambda\mathbf{I})^{1/2})\mathbf{Q}^T\mathbf{L}^\dagger \quad (3)$$

The coefficients $\mathbf{Z}$ update is:

$$\min_{\mathbf{Z}} \|\mathbf{TA} - \mathbf{Z}\|_F^2 + \gamma(\|\mathbf{Z}\|_0) \quad (4)$$

$$\mathbf{Z} = (abs(\mathbf{TA}) \geq \gamma).\mathbf{TA} \quad (5)$$

where the term $\mathbf{TA}$ is hard thresholded against a certain threshold γ and '.' denotes the element-wise product.

The basic transform learning formulation in (1) can be made deep by cascading multiple transforms together to generate the coefficients. The joint optimization formulation for learning a n-layer DTL network is given as [18]:

$$\min_{\mathbf{T_i}'s,\mathbf{Z_n}} \left\|\mathbf{T_n}(\phi(\mathbf{T_{(n-1)}}(..\phi(\mathbf{T_1A})))) - \mathbf{Z_n}\right\|_F^2 + \lambda\sum_{i=1}^{n}(\|\mathbf{T_i}\|_F^2 - \log\det \mathbf{T_i}) \quad (6)$$

where ϕ is the activation function, $\mathbf{T_i}'s$ are the deep transforms for $i = 1,..,n$ and $\mathbf{Z_n}$ is the coefficient of the n^{th}-layer. Here, the coefficients of the 1^{st} layer are fed as input to the 2^{nd} transform layer and so on till the n^{th} transform layer to obtain $\mathbf{Z_n}$. As shown in [18], DTL architectures can learn rich representation from the data and hence can better model the complex data compared to shallow version. Thus, we have employed a DTL-based framework for MISR in our proposed formulation.

III. Joint Coupled Deep Transform Learning (JCDTL) Framework for MISR

The block diagram of the proposed JCDTL framework for MISR is shown in Fig. 1. Let $\mathbf{A}$ represent the HR image of guidance modality and $\mathbf{B}$ and $\mathbf{C}$ represent the LR and HR image of target modality, respectively. n-layer deep transforms are learnt from $\mathbf{A}$ and $\mathbf{B}$ while a single transform is learnt from $\mathbf{C}$. Since the different modalities capture the same scene of interest, their representations (transform coefficients) are related to each other. This relationship is mathematically expressed as: $\mathbf{Z_C} = \mathbf{W_A}\mathbf{Z_{An}} + \mathbf{W_B}\mathbf{Z_{Bn}}$ where $\mathbf{W_A}$, $\mathbf{W_B}$ are the unknown weight matrices, $\mathbf{Z_C}$ is the representation for $\mathbf{C}$ and $\mathbf{Z_{An}}$, $\mathbf{Z_{Bn}}$ are the n^{th} layer DTL coefficients learnt from $\mathbf{A}$ and $\mathbf{B}$ respectively. The JCDTL optimization formulation is given as:

$$\begin{aligned}
&\min_{\substack{\mathbf{T_{Ai}}'s,\mathbf{T_{Bi}}'s,\mathbf{T_C},\mathbf{W_A},\\ \mathbf{W_B},\mathbf{Z_{Ai}}'s,\mathbf{Z_{Bi}}'s,\mathbf{Z_C}}} \left\|\mathbf{T_{An}}(\mathbf{T_{A(n-1)}}(..(\mathbf{T_{A1}}\mathbf{A}))) - \mathbf{Z_{An}}\right\|_F^2 + \\
&\left\|\mathbf{T_{Bn}}(\mathbf{T_{B(n-1)}}(..(\mathbf{T_{B1}}\mathbf{B}))) - \mathbf{Z_{Bn}}\right\|_F^2 + \|\mathbf{T_C}\mathbf{C} - \mathbf{Z_C}\|_F^2 + \\
&\lambda_A\sum_{i=1}^{n}(\|\mathbf{T_{Ai}}\|_F^2 - \log\det \mathbf{T_{Ai}}) + \lambda_B\sum_{i=1}^{n}(\|\mathbf{T_{Bi}}\|_F^2 - \log\det \mathbf{T_{Bi}}) \\
&+\lambda_C(\|\mathbf{T_C}\|_F^2 - \log\det \mathbf{T_C}) + \mu\|\mathbf{Z_C} - \mathbf{W_A}\mathbf{Z_{An}} - \mathbf{W_B}\mathbf{Z_{Bn}}\|_F^2 \\
&+ \gamma(\|\mathbf{Z_{An}}\|_1 + \|\mathbf{Z_{Bn}}\|_1 + \|\mathbf{Z_C}\|_1) \\
&\text{s.t. } \mathbf{T_{A(n-1)}}\,(..(\mathbf{T_{A1}}\mathbf{A})) \geq 0,..,\ \mathbf{T_{A1}}\mathbf{A} \geq 0, \\
&\mathbf{T_{B(n-1)}}(..(\mathbf{T_{B1}}\mathbf{B})) \geq 0,..,\ \mathbf{T_{B1}}\mathbf{B} \geq 0.
\end{aligned} \quad (7)$$

Here for $i = 1,..n$, $\{\mathbf{T_{Ai}}, \mathbf{T_{Bi}}\}$ are the deep transforms and $\{\mathbf{Z_{Ai}}, \mathbf{Z_{Bi}}\}$ are their associated coefficients that are learnt from $\mathbf{A}$ and $\mathbf{B}$. $\mathbf{T_C}$ is the single transform that is learnt from

$\mathbf{C}$ to generate the $\mathbf{Z_C}$. Sparsity is enforced on coefficients $\mathbf{Z_{An}}$, $\mathbf{Z_{Bn}}$ and $\mathbf{Z_C}$ using the l_1- norm constraint. Here, λ_A, λ_B, λ_C, μ and γ are the tunable hyperparameters and a $ReLU$ type non-linearity is considered between the deep layers by forcing the negative values of the coefficients to 0.

In the rest of the section, without loss of generality, we provide the solution steps of the above optimization formulation by assuming $n = 3$. One can derive the solution by following similar steps for any general n. The joint formulation of (7) for $n = 3$ i.e. the 3-layer deep network can be expressed as:

$$\begin{aligned}&\min_{\substack{\mathbf{T_{A1}},\mathbf{T_{A2}},\mathbf{T_{A3}},\mathbf{T_{B1}},\mathbf{T_{B2}},\mathbf{T_{B3}},\mathbf{T_C},\mathbf{W_A},\\ \mathbf{W_B},\mathbf{Z_{A1}},\mathbf{Z_{A2}},\mathbf{Z_{A3}},\mathbf{Z_{B1}},\mathbf{Z_{B2}},\mathbf{Z_{B3}},\mathbf{Z_C}}} (\|\mathbf{T_{A3}Z_{A2}} - \mathbf{Z_{A3}}\|_F^2 + \\ &\|\mathbf{T_{A2}Z_{A1}}-\mathbf{Z_{A2}}\|_F^2+\|\mathbf{T_{A1}A}-\mathbf{Z_{A1}}\|_F^2)+(\|\mathbf{T_{B3}Z_{B2}}-\mathbf{Z_{B3}}\|_F^2 \\ &+\|\mathbf{T_{B2}Z_{B1}}-\mathbf{Z_{B2}}\|_F^2+\|\mathbf{T_{B1}B}-\mathbf{Z_{B1}}\|_F^2)+\|\mathbf{T_C C} - \mathbf{Z_C}\|_F^2 \\ &+\lambda_A\sum_{i=1}^{3}(\|\mathbf{T_{Ai}}\|_F^2-\log\det\mathbf{T_{Ai}})+\lambda_B\sum_{i=1}^{3}(\|\mathbf{T_{Bi}}\|_F^2-\log\det\mathbf{T_{Bi}}) \\ &+\lambda_C(\|\mathbf{T_C}\|_F^2-\log\det\mathbf{T_C})+\mu\|\mathbf{Z_C}-\mathbf{W_A Z_{A3}}-\mathbf{W_B Z_{B3}}\|_F^2 \\ &+\gamma(\|\mathbf{Z_{A3}}\|_1 \ +\|\mathbf{Z_{B3}}\|_1 \ +\|\mathbf{Z_C}\|_1) \\ &\text{s.t.}\,\mathbf{T_{A2}}(\mathbf{T_{A1}A})\geq 0,\,\mathbf{T_{A1}A}\geq 0,\,\mathbf{T_{B2}}(\mathbf{T_{B1}B})\geq 0,\,\mathbf{T_{B1}B}\geq 0.\end{aligned} \tag{8}$$

Here, $\{\mathbf{A},\mathbf{B},\mathbf{C}\} \in \mathbb{R}^{F\times M}$ represent the vectorized 2D image patches of length M with F features (pixels). In our case, the transforms and the weight matrices, $\{\mathbf{T_{A1}},\mathbf{T_{A2}},\mathbf{T_{A3}},\mathbf{T_{B1}},\mathbf{T_{B2}},\mathbf{T_{B3}},\mathbf{T_C},\mathbf{W_A},\mathbf{W_B}\}\in\mathbb{R}^{F\times F}$ and coefficients $\{\mathbf{Z_{A1}},\mathbf{Z_{A2}},\mathbf{Z_{A3}},\mathbf{Z_{B1}},\mathbf{Z_{B2}},\mathbf{Z_{B3}},\mathbf{Z_C}\} \in \mathbb{R}^{F\times M}$. This method involves a training phase where the transforms are learned for the different modalities. Later, they are utilized to generate the HR image of the target modality during the test phase.

Training Phase: In this phase, we employ ADMM based variable splitting approach to compute the closed-form updates for the transforms, corresponding coefficients and weight matrices. The sub-problems to solve for updating the deep transforms for the modality $\mathbf{A}$ are given as:

$$\min_{\mathbf{T_{A1}}}\|\mathbf{T_{A1}A}-\mathbf{Z_{A1}}\|_F^2+\lambda_A(\|\mathbf{T_{A1}}\|_F^2-\log\det\mathbf{T_{A1}}) \tag{9}$$

$$\min_{\mathbf{T_{A2}}}\|\mathbf{T_{A2}Z_{A1}}-\mathbf{Z_{A2}}\|_F^2+\lambda_A(\|\mathbf{T_{A2}}\|_F^2-\log\det\mathbf{T_{A2}}) \tag{10}$$

$$\min_{\mathbf{T_{A3}}}\|\mathbf{T_{A3}Z_{A2}}-\mathbf{Z_{A3}}\|_F^2+\lambda_A(\|\mathbf{T_{A3}}\|_F^2-\log\det\mathbf{T_{A3}}). \tag{11}$$

In the similar way, the sub-problems to solve for updating the deep transforms for modality $\mathbf{B}$ can be obtained by replacing $\mathbf{A}$ with $\mathbf{B}$ in the above (9) - (11). Now, the sub-problem to solve for the transform $\mathbf{T_C}$ is given as:

$$\min_{\mathbf{T_C}}\|\mathbf{T_C C}-\mathbf{Z_C}\|_F^2+\lambda_C(\|\mathbf{T_C}\|_F^2-\log\det\mathbf{T_C}). \tag{12}$$

Notice that (9) - (12) resembles (2) and hence the closed form expressions similar to (3) for updating the transforms can be obtained as described in Sec. II.

The coefficients $\mathbf{Z_{A1}},\mathbf{Z_{A2}}$ for modality $\mathbf{A}$ are computed by solving the sub-problems given below.

$$\begin{aligned}&\min_{\mathbf{Z_{A1}}}\|\mathbf{T_{A2}Z_{A1}}-\mathbf{Z_{A2}}\|_F^2+\|\mathbf{T_{A1}A}-\mathbf{Z_{A1}}\|_F^2 \\ &\text{subject to}\ \mathbf{Z_{A_1}}\geq 0.\end{aligned} \tag{13}$$

$$\begin{aligned}&\min_{\mathbf{Z_{A2}}}\|\mathbf{T_{A3}Z_{A2}}-\mathbf{Z_{A3}}\|_F^2+\|\mathbf{T_{A2}Z_{A1}}-\mathbf{Z_{A2}}\|_F^2 \\ &\text{subject to}\ \mathbf{Z_{A2}}\geq 0.\end{aligned} \tag{14}$$

The closed-form solutions for $\mathbf{Z_{A1}},\mathbf{Z_{A2}}$ are obtained by taking a derivative of the sub-problems with respect to the argument variable and equating it to 0. This results in the following updates:

$$\mathbf{Z_{A1}}=\max(0,(\mathbf{I}+{\mathbf{T_{A2}}}^T\mathbf{T_{A2}})^\dagger.({\mathbf{T_{A2}}}^T\mathbf{Z_{A2}}+\mathbf{T_{A1}A})) \tag{15}$$

$$\mathbf{Z_{A2}}=\max(0,(\mathbf{I}+{\mathbf{T_{A3}}}^T\mathbf{T_{A3}})^\dagger.({\mathbf{T_{A3}}}^T\mathbf{Z_{A3}}+\mathbf{T_{A2}Z_{A1}})) \tag{16}$$

where $\max(.)$ is a greedy approach considered for $ReLU$ type non-linearity for the deep layers. Similarly, the updates for the coefficient $\mathbf{Z_{B1}},\mathbf{Z_{B2}}$ for modality $\mathbf{B}$ can be obtained by replacing $\mathbf{A}$ with $\mathbf{B}$ in (15 and 16). The coefficients of the last layer i.e., $n = 3$ in this case is estimated by solving the following:

$$\begin{aligned}&\min_{\mathbf{Z_{A3}}}\|\mathbf{T_{A3}Z_{A2}}-\mathbf{Z_{A3}}\|_F^2+\gamma\|\mathbf{Z_{A3}}\|_1 \\ &+\mu(\|\mathbf{Z_C}-\mathbf{W_A Z_{A3}}-\mathbf{W_B Z_{B3}}\|_F^2).\end{aligned} \tag{17}$$

Due to the l_1-norm constraint on $\mathbf{Z_{A3}}$, basic matrix manipulation and soft thresholding is used similar to the work in [17] to obtain the closed form update:

$$\mathbf{Z_{A3}}=\text{sign}(\mathbf{X_A}).\max(0,|\mathbf{X_A}|-\mathbf{Y_A}) \tag{18}$$

where $\mathbf{X_A}$ = $\mathbf{D}^\dagger.(\mu{\mathbf{W_A}}^T(\mathbf{Z_C}-\mathbf{W_B Z_{B3}})+\mathbf{T_{A3}Z_{A2}})$, $\mathbf{Y_A}$ = $\mathbf{D}^\dagger.(\frac{\gamma}{2}\mathbf{J})$ and $\mathbf{D}$ = $\mathbf{I}$ + $\mu{\mathbf{W_A}}^T\mathbf{W_A}$, $\mathbf{J}$ is an all ones matrix. In the similar way, the coefficients $\mathbf{Z_{B3}}$ associated with modality $\mathbf{B}$ are updated with the following closed-form update:

$$\mathbf{Z_{B3}}=\text{sign}(\mathbf{X_B}).\max(0,|\mathbf{X_B}|-\mathbf{Y_B}) \tag{19}$$

where $\mathbf{X_B}$ = $\mathbf{E}^\dagger.(\mu{\mathbf{W_B}}^T(\mathbf{Z_C}-\mathbf{W_A Z_{A3}})+\mathbf{T_{B3}Z_{B2}})$, $\mathbf{Y_B}$ = $\mathbf{E}^\dagger.(\frac{\gamma}{2}\mathbf{J})$ and $\mathbf{E}$ = $\mathbf{I}$ + $\mu{\mathbf{W_B}}^T\mathbf{W_B}$.

The sub-problem and the associated closed form solution for $\mathbf{Z_C}$ is given as:

$$\begin{aligned}&\min_{\mathbf{Z_C}}\|\mathbf{T_C C}-\mathbf{Z_C}\|_F^2+\gamma\|\mathbf{Z_C}\|_1 \\ &+\mu(\|\mathbf{Z_C}-\mathbf{W_A Z_{A3}}-\mathbf{W_B Z_{B3}}\|_F^2)\end{aligned} \tag{20}$$

$$\mathbf{Z_C}=\text{sign}(\mathbf{X_C}).\max(0,|\mathbf{X_C}|-\mathbf{Y_C}) \tag{21}$$

where $\mathbf{X_C}=\frac{1}{(1+\mu)}.(\mu(\mathbf{W_A Z_{A3}}+\mathbf{W_B Z_{B3}})+\mathbf{T_C C})$ and $\mathbf{Y_C}=\frac{\gamma}{2(1+\mu)}$.

The weight matrices, $\mathbf{W_A}$ and $\mathbf{W_B}$ are updated by solving the following sub-problems:

$$\min_{\mathbf{W_A}}\|\mathbf{Z_C}-\mathbf{W_A Z_{A_3}}-\mathbf{W_B Z_{B_3}}\|_F^2 \tag{22}$$

$$\min_{\mathbf{W_B}}\|\mathbf{Z_C}-\mathbf{W_A Z_{A_3}}-\mathbf{W_B Z_{B_3}}\|_F^2. \tag{23}$$

The closed form updates for $\mathbf{W_A}$ and $\mathbf{W_B}$ is obtained using simple least squares given as:

$$\mathbf{W_A}=(\mathbf{Z_C}-\mathbf{W_B Z_{B3}})\mathbf{Z_{A3}}^\dagger \tag{24}$$

$$\mathbf{W_B}=(\mathbf{Z_C}-\mathbf{W_A Z_{A3}})\mathbf{Z_{B3}}^\dagger \tag{25}$$

The transforms and coefficients update go through many iterations until convergence is met. This completes the training phase.

Test Phase: In this phase, the learnt transforms for the different modalities and the weight matrices are utilized to compute the coefficients for the test data $\mathbf{A^{test}}$ and $\mathbf{B^{test}}$ later

TABLE I: RGB/NIR dataset and RGB/Multispectral dataset evaluated by PSNR (dB) and SSIM for 16× and 4× upscaling factor respectively. The highest value are highlighted using bold and the second highest value is underlined.

	RGB/NIR Dataset (16× Upsampling)										**RGB/Multispectral Dataset (4× Upsampling)**									
	Indoor 4		Indoor 5		Indoor 11		Indoor 16		Indoor21		Imge6		Imge7		Imgf5		Imgf7		Imgh3	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Proposed JCDTL (3-layer)	**29.062**	**0.916**	**30.104**	0.938	**28.579**	**0.898**	**32.261**	**0.935**	**28.016**	**0.896**	**31.441**	**0.841**	**35.496**	0.899	**37.522**	**0.947**	**33.339**	0.888	**39.403**	**0.948**
JCTL	28.003	0.895	28.822	0.934	27.086	0.883	29.708	0.925	26.783	0.874	28.793	0.814	32.669	0.889	36.277	0.939	31.964	0.864	37.140	0.941
CDL	26.858	0.902	27.784	0.915	26.332	0.830	29.663	0.891	26.112	0.850	31.049	0.835	33.222	0.877	34.239	0.906	31.401	0.878	36.107	0.920
DL	25.998	0.848	26.994	0.878	26.010	0.838	29.558	0.860	25.212	0.830	20.968	0.828	26.732	**0.938**	32.588	0.824	23.851	**0.902**	30.788	0.924
JR	23.465	0.862	26.407	**0.949**	23.137	0.828	23.793	0.872	22.066	0.809	26.519	0.814	32.781	0.889	33.933	0.890	29.295	0.804	33.999	0.922
GF	24.779	0.890	24.654	0.876	23.868	0.745	28.715	0.892	23.661	0.780	25.332	0.774	29.709	0.869	31.706	0.901	28.045	0.880	33.518	0.807
JBF	23.710	0.853	25.422	0.889	24.185	0.805	28.395	0.896	23.605	0.814	25.535	0.746	29.655	0.799	32.411	0.886	28.874	0.839	34.461	0.902

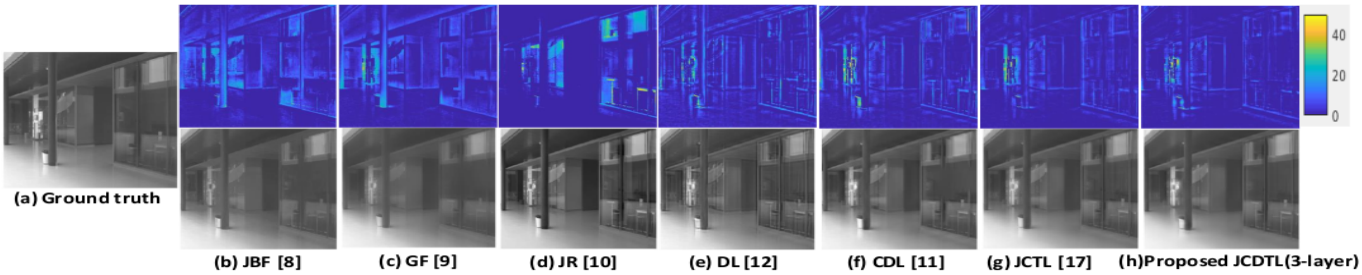

Fig. 2: Visual comparison for RGB/NIR for Indoor16. The top row describes the error map and the bottom row has the reconstructed image for different methods

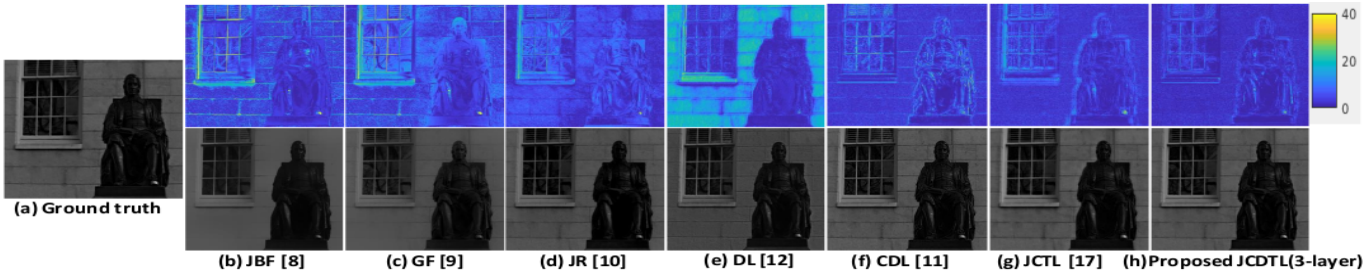

Fig. 3: Visual comparison for RGB/Multispectral of 640nm band for Imge7. The top row describes the error map and the bottom row has the reconstructed image for different methods

estimate $\mathbf{C^{test}}$. The test coefficients of the first two layers for modality $\mathbf{A}$ are computed as: $\mathbf{Z^{test}_{A1}} = \mathbf{T_{A1}A^{test}}$ and $\mathbf{Z^{test}_{A2}} = \mathbf{T_{A2}}\ \mathbf{Z^{test}_{A1}}$. Similarly, the test coefficients of the first two layers for modality $\mathbf{B}$ are given as: $\mathbf{Z^{test}_{B1}} = \mathbf{T_{B1}}\ \mathbf{B}$ and $\mathbf{Z^{test}_{B2}} = \mathbf{T_{B2}}\ \mathbf{Z^{test}_{B1}}$. For the update of $\mathbf{Z^{test}_{A3}}$ and $\mathbf{Z^{test}_{B3}}$ the following sub-problems need to be solved:

$$\min_{\mathbf{Z^{test}_{A3}}} \|\mathbf{T_{A3}Z^{test}_{A2}} - \mathbf{Z^{test}_{A3}}\|_F^2 + \gamma \left\|\mathbf{Z^{test}_{A3}}\right\|_1 \quad (26)$$

$$\min_{\mathbf{Z^{test}_{B3}}} \|\mathbf{T_{B3}Z^{test}_{B2}} - \mathbf{Z^{test}_{B3}}\|_F^2 + \gamma \left\|\mathbf{Z^{test}_{B3}}\right\|_1. \quad (27)$$

These are standard expression for LASSO based optimization problems [25] for which the closed form update is given as:

$$\mathbf{Z^{test}_{A3}} = \text{sign}(\mathbf{T_{A3}Z^{test}_{A2}}).\max(0, |\mathbf{T_{A3}Z^{test}_{A2}}| - \frac{\gamma}{2}) \quad (28)$$

$$\mathbf{Z^{test}_{B3}} = \text{sign}(\mathbf{T_{B3}Z^{test}_{B2}}).\max(0, |\mathbf{T_{B3}Z^{test}_{B2}}| - \frac{\gamma}{2}). \quad (29)$$

Subsequently, $\mathbf{Z^{test}_C}$ and reconstructed image $\mathbf{C^{test}}$ is computed as: $\mathbf{Z^{test}_C}$= $\mathbf{W_A Z_{A3}}$+$\mathbf{W_B Z_{B3}}$ and $\mathbf{C_{test}}$=$\mathbf{T^{\dagger}_C}\ \mathbf{Z^{test}_C}$.

IV. Results and Discussions

The performance of the proposed method is evaluated on two different multimodal datasets; namely, RGB-Multispectral dataset [20] and RGB-NIR dataset [19]. The RGB image is considered as the guidance modality in both datasets, and the Multispectral/NIR image as the target modality. For the RGB-Multispectral dataset, the image data at 640 nm is considered. The results of the proposed approach are compared against six state-of-the-art MISR techniques based on different methods namely, dictionary learning (CDL [11]), deep learning (DL [12]), guided image filtering (GF [9]), joint bilateral filtering (JBF [8]), joint image restoration (JR [10]) and the shallow variant of the proposed technique (JCTL [17]). Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR) metrics are evaluated to estimate the reconstruction quality of the HR image of the target modality.

The two datasets considered here contain the HR images of both the guidance ($\mathbf{A}$) and target modalities ($\mathbf{C}$). Similar to [11], the LR image of target modality ($\mathbf{B}$) for both the datasets, is generated by downsampling $\mathbf{C}$ by a required factor and then applying bicubic interpolation on the downsampled image and upscaling by the same factor. For comparison with other techniques, the RGB/Multispectral data is downsampled by 4×, whereas the RGB/NIR dataset is downsampled by 16× like in [17]. The RGB image used as guidance modality is converted to grayscale. Each image is truncated into patches of size 16×16 and then converted into a vectorized patch. In the training phase, patches of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are chosen to learn model parameters. The hyperparameters λ_A,λ_B,λ_C, μ and γ were chosen using grid search.

The PSNR and SSIM reconstruction results obtained with both datasets namely RGB/NIR and RGB/Multispectral are tabulated in Table 1 for 5 test images. It can be observed

that the proposed JCDTL method with 3-layers performs best in terms of PSNR and for most cases in terms of SSIM on both datasets compared to other methods. One can also observe a considerable improvement close to 3dB (for e.g., see the results for imge 6 and imge 7) in some cases with the proposed JCDTL compared to JCTL. The hyperparameters used for the proposed 3-layer JCDTL method for RGB/NIR dataset are λ_A=λ_B=λ_C=2.8, μ=2.1e-4 and γ=4e-3, while for RGB/Multispectral, they are λ_A=λ_B=λ_C=4.4, μ=1e-4 and γ=4e-4. For visual comparison, Fig. 2 and Fig. 3 presents the reconstructed image obtained with different techniques on the test image on both datasets. The bottom row shows the reconstructed image, while the top row provides the corresponding error map. A considerable improvement in the error map with the proposed JCDTL approach can be noticed from these figures compared to other approaches.

Next, we provide the results to illustrate the effect of increase in the layers. Fig. 4 shows the plot of PSNR vs. the number of layers for both the image datasets (averaged over the five test images). Observe from the plot that a noticeable improvement can be seen by increasing the layers from 1 to 3, however the improvement seems to diminish beyond layer 3. This plot reflects the law of diminishing returns and hence in our work we considered a 3-layer JCDTL for target image reconstruction. Also, notice from the graph that compared to the shallow version (JCTL), the 3-layer JCDTL gave an average PSNR improvement of around 1.5dB and 2dB for RGB/NIR and RGB/Multispectral data, respectively.

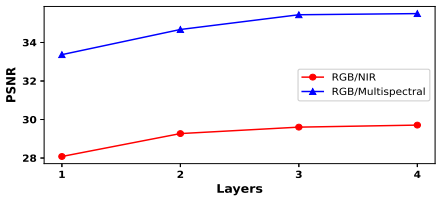

Fig. 4: PSNR performance of JCDTL with different layers on both datasets

V. Conclusion

In this paper, we have introduced a novel DTL based approach referred as JCDTL for MISR. The generic n-layer deep optimization formulation for JCDTL method which exploits the cross-modal dependencies is provided, and the requisite closed-form updates are provided. Results obtained with two publicly available RGB/NIR and RGB/multi-spectral datasets, demonstrate the improved reconstruction performance of the proposed JCDTL compared to state-of-the-art methods. In future, we plan to explore the hybrid deep transform and deep dictionary learning based frameworks to obtain richer representation for improved performance.

References

[1] G. A. Shaw and H. K. Burke, "Spectral imaging for remote sensing," *Lincoln laboratory journal*, vol. 14, no. 1, pp. 3–28, 2003.

[2] A. F. Goetz, "Three decades of hyperspectral remote sensing of the earth: A personal view," *Remote Sensing of Environment*, vol. 113, pp. S5–S16, 2009.

[3] T. Zhang, W. Wei, B. Zhao, R. Wang, M. Li, L. Yang, J. Wang, and Q. Sun, "A reliable methodology for determining seed viability by using hyperspectral data from two sides of wheat seeds," *Sensors*, vol. 18, no. 3, p. 813, 2018.

[4] B. Zhang, D. Wu, L. Zhang, Q. Jiao, and Q. Li, "Application of hyperspectral remote sensing for environment monitoring in mining areas," *Environmental Earth Sciences*, vol. 65, no. 3, pp. 649–658, 2012.

[5] B. Park, K. C. Lawrence, W. R. Windham, D. P. Smith, and P. W. Feldner, "Hyperspectral imaging for food processing automation," in *Imaging Spectrometry VIII*, vol. 4816. International Society for Optics and Photonics, 2002, pp. 308–316.

[6] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of biomedical optics*, vol. 19, no. 1, p. 010901, 2014.

[7] D. B. Malkoff and W. R. Oliver, "Hyperspectral imaging applied to forensic medicine," in *Spectral Imaging: Instrumentation, Applications, and Analysis*, vol. 3920. International Society for Optics and Photonics, 2000, pp. 108–116.

[8] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics (ToG)*, vol. 26, no. 3, pp. 96–es, 2007.

[9] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 1397–1409, 06 2013.

[10] X. Shen, Q. Yan, L. Xu, L. Ma, and J. Jia, "Multispectral joint image restoration via optimizing a scale map," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 12, pp. 2518–2530, 2015.

[11] P. Song, X. Deng, J. F. Mota, N. Deligiannis, P. L. Dragotti, and M. R. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 57–72, 2019.

[12] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *European Conference on Computer Vision*. Springer, 2016, pp. 154–169.

[13] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof, "A deep primal-dual network for guided depth super-resolution," *arXiv preprint arXiv:1607.08569*, 2016.

[14] X. Song, Y. Dai, and X. 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.

[15] X. Deng and P. L. Dragotti, "Coupled ista network for multi-modal image super-resolution," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1862–1866.

[16] S. Ravishankar and Y. Bresler, "Learning overcomplete sparsifying transforms for signal processing," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3088–3092.

[17] A. Gigie, A. A. Kumar, A. Majumdar, K. Kumar, and M. G. Chandra, "Joint coupled transform learning framework for multimodal image super-resolution," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1640–1644.

[18] J. Maggu and A. Majumdar, "Greedy deep transform learning," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1822–1826.

[19] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *CVPR 2011*. IEEE, 2011, pp. 177–184.

[20] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *CVPR 2011*. IEEE, 2011, pp. 193–200.

[21] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2013.

[22] J. Maggu and A. Majumdar, "Unsupervised deep transform learning," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6782–6786.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. 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.

[24] S. Ravishankar and Y. Bresler, "Closed-form solutions within sparsifying transform learning," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 5378–5382.

[25] N. Gauraha, "Introduction to the lasso," *Resonance*, vol. 23, no. 4, pp. 439–464, 2018.