BUSIFusion: Blind Unsupervised Single Image Fusion of Hyperspectral and RGB Images

Jiabao Li, Yuqi Li *Member, IEEE,* Chong Wang, Xulun Ye, and Wolfgang Heidrich *Fellow, IEEE*

***Abstract*—Hyperspectral images (HSIs) provide rich spectral information that has been widely used in numerous computer vision tasks. However, their low spatial resolution often prevents their use in applications such as image segmentation and recognition. Fusing low-resolution HSIs with high-resolution RGB images to reconstruct high-resolution HSIs has attracted great research attention recently. In this paper, we propose an unsupervised blind fusion network that operates on a single HSI and RGB image pair and requires neither known degradation models nor any training data. Our method takes full advantage of an unrolling network and coordinate encoding to provide a state-of-the-art HSI reconstruction. It can also estimate the degradation parameters relatively accurately through the neural representation and implicit regularization of the degradation model. The experimental results demonstrate the effectiveness of our method both in simulations and in our real experiments. The proposed method outperforms other state-of-the-art nonblind and blind fusion methods on two popular HSI datasets. Our related code and data is available at https://github.com/CPREgroup/Real-Spec-RGB-Fusion.**

***Index Terms*—Unsupervised Image Fusion, Blind Fusion, Hyperspectral Image Fusion**

I. Introduction

Hyperspectral images (HSIs) are three-dimensional data cubes with two spatial dimensions and a spectral dimension. Since HSIs contain rich spectral signatures that can describe physical properties and chemical composition of objects, HSIs are widely used in diverse applications such as land cover classification [1], biological recognition [2], and high-color-fidelity display [3]. To acquire high-quality HSIs for the demands of these applications, hyperspectral imaging has gathered substantial attention from the computer vision community in the past two decades [4]–[6].

Although HSI cameras can directly capture images with tens to hundreds of spectral bands, they either suffer from relatively low spatial resolution and the narrow depth-of-field, or require temporal scanning in either the spatial or spectral dimension. Either of these scenarios presents a strong obstacle to applications such as object tracking [7], [8], image classification [9]–[12] and segmentation [13]. One of the commonly used methods to practically reconstruct high-resolution hyperspectral images is fusing a low-resolution hyperspectral image captured by a hyperspectral sensor with a corresponding high-resolution RGB image captured by a conventional RGB sensor [14]. However, HSI and RGB image fusion is an ill-posed problem since both the high-resolution RGB and the low-resolution HSI are heavily degraded from the high-resolution HSI with the spectral sensitivity/response functions (SSFs or SRFs) and the point spread function (PSF) kernels respectively [15]–[17].

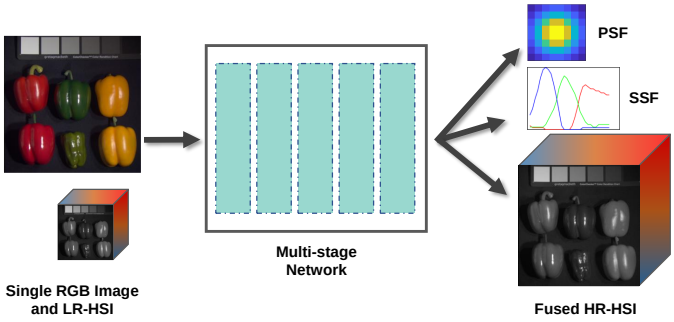


Fig. 1. The framework of our *unsupervised single* HSI and RGB image *blind* fusion method.

Most of the previous fusion works [18], [19], assume that both SSFs and PSFs are available and known [20]–[22], and recover high-resolution HSIs by either solving the inverse problem of the known degradation model or learning the reconstruction from the dataset synthesized using the SSFs and PSFs. However, the assumption is impractical since the measurements of SSFs and PSFs usually require professional devices such as spectrometer and ideal point light source. Furthermore, dataset synthesis methods of previous works are unrealistic because they apply a single PSF kernel to generate low-resolution HSIs in the entire dataset, whereas real PSF kernels depend on the spectral composition of the scene and vary with scene depth.

Therefore, to avoid acquiring SSFs and PSFs, and to ensure adopting scene-variant PSFs, in this paper, we propose an *unsupervised snapshot* hyperspectral and RGB image *blind* fusion method, which jointly recovers an high-resolution HSI and degradation parameters (SSFs and PSFs) only from a single pair of a low-resolution HSI and an aligned high-resolution RGB image without any known imaging degradation parameters or constructed dataset, as shown in Fig.1. Compared to the previous methods with either known degradation parameters or a large-scale dataset, our problem is the most challenging due

The work has been supported by the Natural Science Foundation of Zhejiang Province(Y19F020050). (Corresponding author: Yuqi Li.)

Jiabao Li, Yuqi Li, Chong Wang, Xulun Ye are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315600, China (e-mail: 2011082305@nbu.edu.cn; liyuqi1@nbu.edu.cn; wangchong@nbu.edu.cn; yexulun@nbu.edu.cn). Yuqi Li is also with Zhejiang Engineering Research Center of Advanced Mass Spectrometry and Clinical Application.

Wolfgang Heidrich is with the Visual Computing Center, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia (e-mail: wolfgang.heidrich@kaust.edu.sa).

Manuscript received April 19, 2022; revised August 16, 2022.

TABLE I
CLASSIFICATION OF HSI AND MSI FUSION METHODS DEPENDING ON THE PRESENCE OF THE DEGRADATION PARAMETERS (PSF AND SSF) AND THE TRAINING DATA.

Types	Descriptions	Methods
Unsupervised & Blind	Unknown PSF & SSF w/o training set	[23]–[27]
Unsupervised & Semi-blind	Unknown PSF or SSF w/o training set	[28]–[30]
Unsupervised & Nonblind	Known PSF & SSF w/o training set	[31]–[49]
Supervised & Blind	Unknown PSF & SSF with training set	[50]–[54]
Supervised & Nonblind	Known PSF & SSF with training set	[55]–[58]

to the larger degrees of optimization freedom, while requiring the fewest known conditions.

The proposed method, named *BUSI-FusionNet*, reconstructs both the high-resolution HSI and the degradation parameters through deep unsupervised learning based on a physical model. Our blind method can even achieve more accurate fusion than previous nonblind methods by utilizing deep image denoising prior of a single image and neural representation of degradation parameters. The contributions of this paper can be summarized as follows:

- We present an unsupervised learning method to jointly recover high-resolution HSIs and the imaging degradation parameters **only** from a single pair of low-resolution HSI and high-resolution RGB image, **without** known SSFs and PSF, or constructing a training set. Our unsupervised blind fusion approach outperforms state-of-the-art unsupervised semi-blind/blind methods.
- We show that the physics-model-based unrolling architecture combined with spatial variant denoising blocks and the implicit neural representations of imaging degradation parameters can guarantee both the data fidelity and the deep prior of the target image.
- We build a high-quality real low-resolution HSI and high-resolution RGB image datasets, providing a general-purpose benchmark for the training and evaluation of real unregistered HSI fusion task.

II. RELATED WORK

In the last decade, many models and optimization approaches have been developed for HSI fusion tasks [59], including pan-sharpening models [60], [61], matrix factorization based models [23], [34], [37], [44], [62], tensor representation based models [33], [36], [63]–[65], and deep learning based models [27]–[30], [38], [47]–[52], [66]–[73]. Most of the models utilize either the prior knowledge of the degradation model or sufficient supervised training data, as shown in Table I.

a) Non-blind methods: Non-blind methods [33], [37], [41] use a known and calibrated physical model to guide the optimization of HSIs fusion. However, since HSI fusion is an ill-posed problem, to reduce the number of possible solutions, some priors are proposed to enhance the robustness of the fusion methods. For example, low-rank regularizers are utilized in unmixing-based methods [37], [42], [43] to constrain the number of final spectra; dictionary-learning-based methods learn a spectral dictionary from the low-resolution HSI, map the spectral dictionary to RGB dictionary using known SSFs, and apply sparse regularizers to obtain the HR HSIs [35], [44], [45], [74]; Bayesian-based method [34] assumes that the representation coefficients of high-resolution HSI follow a Gaussian distribution, and use the assumption as a prior to accomplish high-resolution HSI fusion; tensor-based methods [32], [36], [38] exploit the redundancy of HSI, group non-local similar cube patches to aggregate tensors, and apply sparse representation to construct the tensors. Especially, Liu *et al.* [31] recast the tensor-trace-norm formulation to reconstruct HR HSIs via low-rank approximation; Dian *et al.* [39] use low tensor-train rank (LTTR) as a regularization term on the grouped tensors consisting of non-local patches. Similarly, Xu *et al.* [40] enforces that RGB image and low-resolution HSI share the same factor matrices in the Canonical Polyadic (CP) decomposition of the non-local tensor.

The advent of deep-learning-based methods shows that HSI fusion can achieve superior performance and high efficiency through the use of deep image priors. The deep neural networks [46], [55] adopt convolutional layers to generate the high-resolution HSIs which satisfy the known degradation model, and give better results compared with the methods using handcrafted priors. Others [47], [48] apply cross-space attention blocks and learn a Dirichlet distribution respectively to further improve the fusion quality.

b) Supervised methods: supervised methods [56]–[58] do not reconstruct high-resolution HSI with a known degradation model, instead, they construct the training set of HSIs and RGB images using hyperspectral and RGB cameras, and train neural networks to learn the fusion for the specific devices. Recently some supervised methods are proposed to jointly learn the degradation model and the fusion from the training set, and iteratively improve the reconstruction results by the multi-blocks neural networks [50], [52]–[54]. However, these supervised methods rely on large-scale high-quality training sets, otherwise, the training sets are synthesized using the ground truth of degradation parameters. Both of the prior knowledge is impractical to obtain, in order to fully address the fusion with the lack of imaging model and perfect training data, the unsupervised blind fusion single image methods are still sorely needed.

c) Unsupervised blind methods: Unsupervised blind methods attempt to fuse a spatially degraded low-resolution HSI and a spectrally degraded high-resolution RGB image to obtain the target high-resolution HSI without a known degradation model and training set. The most straightforward solution is estimating SSF and PSF first, and applying non-blind reconstruction method to recover the target high-resolution HSIs [25]. However, such two steps strategy can

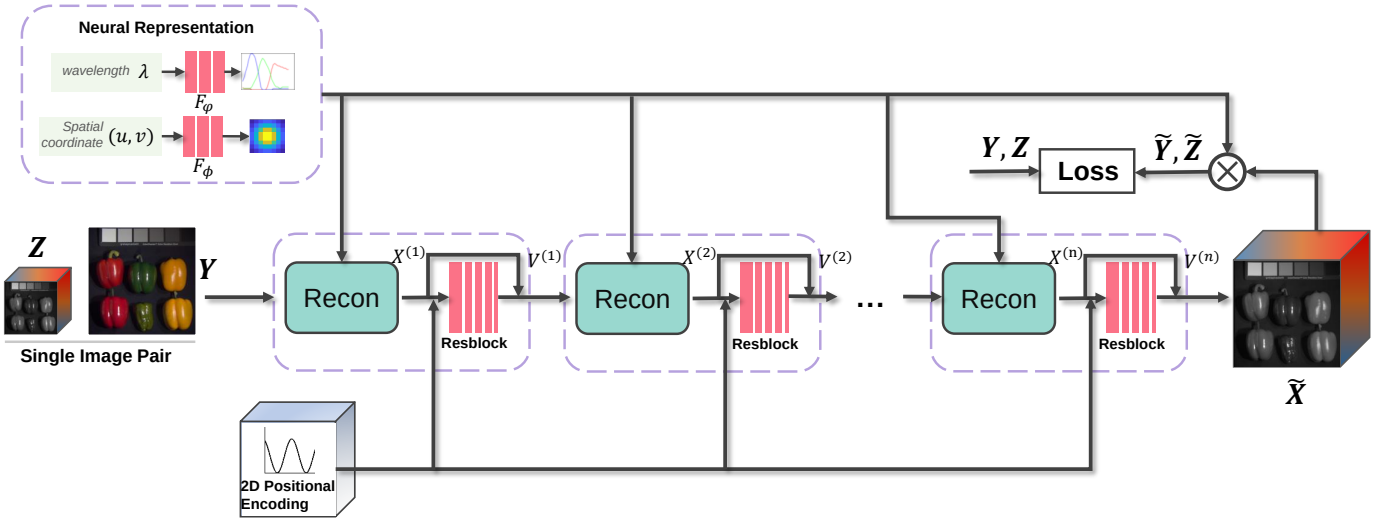


Fig. 2. Architecture of the proposed unsupervised single image blind fusion network. The network is individually trained for each dataset, where both SSFs and PSFs are represented by MLPs for the reconstruction process. To handle the metamerism problem, the tensors sent to each resblock are constructed by concatenating a 2D Positional Encoding tensor with the data tensor X. We use the projection error at the neural representation of SSF and PSF as the loss function to train the unrolling network.

not guarantee the optimal solution of both degradation parameters and the reconstructed images. Therefore, most of the methods jointly estimate the degradation parameters and the high-resolution HSI, which are highly under-constrained. To deal with the highly ill-posed problem, they apply total variation (TV) [23] and L2 norm [24] as regularizers in the loss functions to robustly estimate PSFs and SSFs. The main disadvantage of these methods is the need to adjust the weights of the explicit regularizers in loss functions to adapt to the numeric magnitude of the reprojection error for each image. Others implicitly model the degradation operation in deep networks. For instance, Zheng *et al.* [28] proposed a deep network consisting of three coupled autoencoders, where HSI is unmixed and degradation parameters are adaptively learned; Yao *et al.* [29] propose a coupled convolutional network, where a cross-attention module is embedded to extract and transfer spectral or spatial information.

III. Method

A. Overview

Our goal is to generate an high-resolution HSI from the acquired high-resolution RGB and low-resolution HSI. Consider an high-resolution HSI $\mathbf{X} \in \mathcal{R}^{mn \times k}$, where mn denotes the spatial resolution of the image, and k denotes the spectral resolution. The acquisition of the corresponding high-resolution RGB image $\mathbf{Y} \in \mathcal{R}^{(mn)\times 3}$ follows the degradation model:

$$\mathbf{Y} = \mathbf{X}\mathbf{S} \tag{1}$$

where $\mathbf{S} \in \mathcal{R}^{k\times 3}$ is the SSFs of the RGB camera. The degradation model of the low-resolution HSI $\mathbf{Z} \in \mathcal{R}^{m'n'\times k}$ is constructed as:

$$\mathbf{Z} = \mathbf{\Phi}\mathbf{C}\mathbf{X} \tag{2}$$

where $\mathbf{C} \in \mathcal{R}^{mn\times mn}$ denotes a PSF convolutional operation, which is formulated as a matrix here, and $\mathbf{\Phi} \in \mathcal{R}^{m'n'\times mn}$ denotes a downsampling matrix that uniformly samples the original data by a downsampling ratio r ($m' = m/r, n' = n/r$). We let $\mathbf{C}_\Phi = \mathbf{\Phi}\mathbf{C}$ for abbreviation. The above two degradation models can be regarded as two linear low-dimensional projections of the target high-resolution HSI.

To recover the target high-resolution HSI from a single pair of high-resolution RGB and low-resolution HSI without knowing SSF and PSF, we adopt a neural network architecture composed of multiple unrolling optimization stages. As shown in Fig.2, each stage is constructed via unrolling our optimization iterations for HSI fusion. Note that both the SSF matrix $\mathbf{S}$ and the downsampling matrix $\mathbf{C}_\Phi$ are treated as learned parameters in our network.

B. Unrolling Network

To present the network architecture, we first mathematically formulate the fusion problem as an unconstrained optimization problem, and then loop-unroll the optimization to construct our multi-stage unrolling network.

Our optimization model aims at minimizing the weighted sum of the projection errors of the reconstructed high-resolution HSI. The objective function is formulated as:

$$\arg\min_{\mathbf{X}} \frac{1}{2}||\mathbf{Y} - \mathbf{X}\mathbf{S}||_F^2 + \frac{\eta}{2}||\mathbf{Z} - \mathbf{C}_\Phi\mathbf{X}||_F^2 + J(\mathbf{X}), \tag{3}$$

where $||\cdot||_F$ denotes the Frobenius norm, η denotes a weight for the tradeoff of the high-resolution RGB projection error and the low-resolution HSI projection error, and $J(\cdot)$ denotes the denoising prior of the target high-resolution HSI for regularization. Since $J(\cdot)$ is not differentiable, we introduce an auxiliary variable $\mathbf{V}$ and represent Eq.3 as a constrained optimization problem:

$$\arg\min_{\mathbf{X}} \frac{1}{2}||\mathbf{Y}-\mathbf{X}\mathbf{S}||_F^2+\frac{\eta}{2}||\mathbf{Z}-\mathbf{C}_\Phi\mathbf{X}||_F^2+J(\mathbf{V}), \quad st.\ \mathbf{X} = \mathbf{V}. \tag{4}$$

In order to solve the above optimization problem, we apply half-quadratic-splitting (HQS) method to separate it into two sub-problems respect to $\mathbf{X}$ and $\mathbf{V}$, and alternatively optimize

the two sub-problems with multiple iterations. The update steps of the i-th iteration are formulated as:

$$\begin{cases} \mathbf{X}^{(i+1)} \leftarrow \arg\min_{\mathbf{X}} \frac{1}{2}||\mathbf{Y}-\mathbf{X}\mathbf{S}||_F^2 + \frac{\eta}{2}||\mathbf{Z}-\mathbf{C}_\Phi\mathbf{X}||_F^2 \\ \qquad\qquad + \frac{\beta^{(i+1)}}{2}||\mathbf{X}-\mathbf{V}^{(i)}||_F^2, \\ \mathbf{V}^{(i+1)} \leftarrow \arg\min_{\mathbf{Z}} \frac{\beta^{(i+1)}}{2}||\mathbf{X}^{(i+1)}-\mathbf{V}||_F^2 + J(\mathbf{V}), \end{cases} \tag{5}$$

where $\beta^{(i+1)}$ is a trainable weight term in the i-th iteration.

Optimization of X: It is evident that the optimization of $\mathbf{X}$ in the first subproblem can be treated as solving a Sylvester equation. However, we observe that the eigen-decomposition involved in the optimization process causes instability in the training. Previous works [75], [76] have also shown that obtaining an exact solution in each step is not necessary and is less flexible compared with the forward gradient descent method. Therefore, we apply the general gradient descent method to solve the first subproblem, the update step of $\mathbf{X}^{(i+1)}$ can be performed as:

$$\begin{aligned} \mathbf{X}^{(i+1)} \leftarrow & \mathbf{X}^{(i)} - \alpha^{(i+1)}\Big((\mathbf{X}^{(i)}\mathbf{S}-\mathbf{Y})\mathbf{S}^T + \eta\mathbf{C}_\Phi^T(\mathbf{C}_\Phi\mathbf{X}^{(i)}-\mathbf{Z}) \\ & + \beta^{(i+1)}(\mathbf{X}^{(i)}-\mathbf{V}^{(i)})\Big), \end{aligned} \tag{6}$$

where $\alpha^{(i+1)}$ is the trainable step length in the $(i+1)$th iteration.

Optimization of V: To optimize the auxiliary variable $\mathbf{V}^{(i+1)}$, supervised unrolling networks [75] typically apply convolutional residual denoising modules, which are designed to be spatially invariant. However, in the unsupervised single HSI fusion task, it is difficult to correctly resolve metamerism issues with spatially invariant reconstruction kernels. Spatial invariance also results in lower degrees of freedom and reduced approximation power of the neural network. Therefore, unlike previous denoising modules, our model not only implicitly restricts that each patch satisfies the unified local prior, but also introduces spatially-variant to the features of different pixels to avoid underfitting. The most straightforward idea of introducing spatial-variant is utilizing position-related information. Instead of directly using the coordinates, we apply positional encoding [77] to map the single coordinate value of a pixel into a higher $(2L+1)$ dimensional space:

$$\gamma(p) = \Big(p, sin(2^0\pi p), cos(2^0\pi p), \cdots, sin(2^{L-1}\pi p), cos(2^{L-1}\pi p)\Big), \tag{7}$$

where p denotes the normalized coordinate value lying in $[-1,1]$, and L denotes the order of frequency of the vector. We apply the function $\gamma()$ separately to the two coordinate values of the pixels in the target HSI, and construct the $M \times N \times (4L+2)$ tensor Γ. Then we concatenate Γ with $\mathbf{X}^{i+1}$, and feed the concatenated tensor into the denoising module $\mathcal{D}()$ to obtain $\mathbf{V}^{(i+1)}$:

$$\mathbf{V}^{(i+1)} \leftarrow \mathcal{D}\Big(Concat\Big[\mathbf{X}^{(i+1)}, \Gamma\Big]; \theta\Big). \tag{8}$$

Our denoising module consists of two cascaded residual blocks, each containing five convolutional layers with 3×3 kernels.

By unrolling the alternative optimization of $\mathbf{X}$ and $\mathbf{V}$, we construct a multi-stage neural network, in which each stage consists of a linear mapping module to update $\mathbf{X}$ and a convolutional module to update $\mathbf{V}$. The linear module ensure the reconstructed HSI satisfy the linear mapping model in Eq.3 while the convolutional module can efficiently fit the target HSI in a few iterations.

C. Implicit Neural Representation of SSF and kernel

Now we turn to discuss how to deal with unknown $\mathbf{S}$ and $\mathbf{C}$. The naive solution is directly setting them as trainable matrix and optimizing them jointly with other parameters in the network. However, the optimization degree of freedom of this solution is too high, and the training may fall into a local optimum. Fortunately, most convolutional kernels of hyperspectral cameras and SSFs of RGB cameras are smooth due to their optical and material property. Previous works [23], [24] presented that blind image reconstruction tasks can benefit from the smoothing regularization of SSF and convolutional kernel. In this paper, instead of adding handcrafted regularizer in the loss function, we propose adopting multilayer perception (MLP) networks to represent SSF and convolutional kernel. The major advantage of the implicit neural representation is that the implicit regularization is only relevant to the degradation model, and is not affected by different image reconstruction errors in loss function. Furthermore, the optimization freedom of neural representation is higher than that of the handcrafted ones.

As shown in Fig.2, the SSF and PSF are approximated with MLP networks F_φ and F_ϕ respectively. We optimize the weights φ of F_φ to map from each input wavelength λ to its corresponding spectral sensitivity of the RGB channels, and optimize the weights ϕ of F_ϕ to map from each input 2D coordinates (u,v) of the kernel window to its corresponding intensity. Note that we apply the same positional coding as Eq.7 to map λ and (u,v) to higher dimensional space before feeding them to F_φ and F_ϕ respectively. Both networks F_φ and F_ϕ are fully connected networks consisting of five layers each. Since the blind fusion problem is ill-posed, the shape of the estimated PSF and SSF would not be accurate if we treat them as completely unconstrained trainable parameters. The positional encoding gives constraints on the frequency of the PSF and SSF as regularizers. Since most of the PSF and SSF are either unimodal functions or bimodal functions, it is reasonable to apply positional coding to guarantee both smooth constraint and the freedom of representing PSFs and SSFs. In addition, the dimension of positional coding can be adjusted by controlling the frequency order L to avoid both overfitting and underfitting.

D. Training

Once the K-stage network is built, we train it to learn the network parameters θ, ϕ, φ , and $\{\alpha_i, \beta_i | i = 1, ..., K\}$ simultaneously. In order to be consistent with the objective function, our loss function is designed as:

$$\mathcal{L}oss = ||\mathbf{Y}-\widetilde{\mathbf{X}}\mathbf{S}||_F^2 + \eta\,||\mathbf{Z}-\mathbf{C}_\Phi\widetilde{\mathbf{X}}||_F^2, \tag{9}$$

where $\widetilde{\mathbf{X}}$ is the reconstructed HSI of our network.

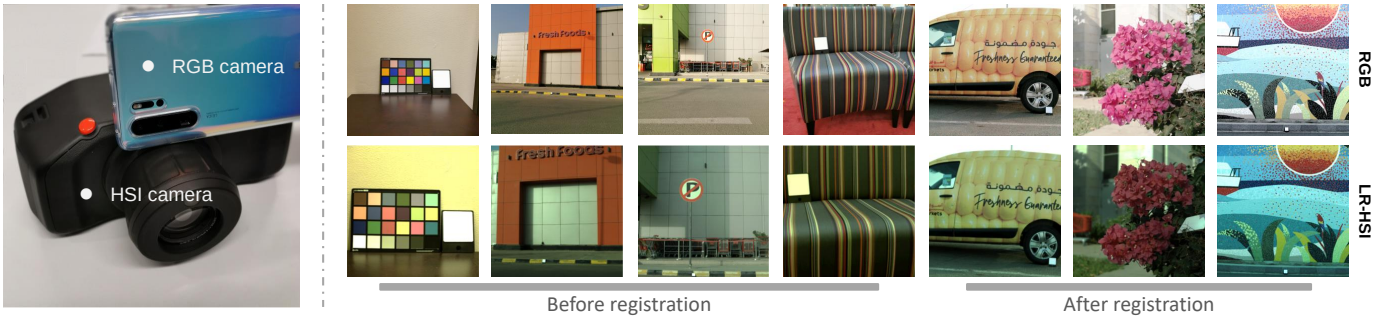


Fig. 3. The capturing devices (left) and our real image dataset (right). Note that the image registration process is applied to align the RGB images and HSI images before sending the image pairs to the proposed BUSI-Fusion network. Here, the optical flow method GMA [78] is adopted for image registration. We use the GMA model on the Sintel dataset [79].

We implement our network in the PyTorch framework, the stage number of K is set to be 6 from experience, the network is trained using an ADAM optimizer with a learning rate of 0.001. We set the parameter $\eta = 1$ for balancing the contribution of the low-resolution HSI and high-resolution RGB image. Trainable variables of the network β and α are empirically initialized to 0.1, note that they should be non-negative so we take them as the input of the ReLU function first in practice. We send the whole RGB image and low-resolution HSI into the network for training. It takes about 200 epochs, which consumes 2.5 hours, to converge and reaches the top quality of the output high-resolution HSI on a V100 GPU.

IV. Experimental

A. Synthetic & Real Datasets

Experiments are conducted on both synthetic datasets, CAVE [80] and KAUST [81], and also on real image pairs that were taken in actual indoor and outdoor scenes. The images in the CAVE[1] and KAUST[2] datasets have the same spatial resolution of 512×512 pixels. We used the first 31 bands of HSI in both datasets, starting from 400nm with 10nm intervals. Ten images and eight images were respectively randomly chosen from CAVE and KAUST for evaluation.

In the ablation study and simulations, the SSFs of Nikon D700 were used to synthesize the RGB images, and two PSFs with size 8×8 and 32×32 were used to generate the low-resolution HSIs. The 32×32 PSF is an average kernel and collocates with 32×32 downsampling operation, and 8×8 PSF is a Gaussian kernel with $\sigma = 2.0$ and collocates with 8×8 downsampling operation.

The real image dataset consists of 200 paired low-resolution HSIs and trichromatic images. The trichromatic images were captured by a HUAWEI P30Pro RYYB camera and a HUAWEI P20 RGGB camera with the spatial resolution 5472×7296, and the HSIs were captured by a compact scanning-based hyperspectral camera Specim IQ with spatial resolution 512×512 and 204 bands ranging from 400nm to 1000nm. We attached the trichromatic camera to the hyperspectral camera as shown in Fig.3 to capture scenes. The extrinsic of the two cameras are close compared to the shooting distance thus the occlusion area can be ignored when we register for the two images. The field of view (FOV) of trichromatic cameras is near three times larger than that of the hyperspectral camera, so we cropped the central region of the trichromatic image to the size 3100×3100 to match the FOV of the hyperspectral camera. To maintain consistency with the synthetic datasets, we uniformly sampled the hyperspectral images in the visible range from 400nm to 700 nm into 10 nm intervals, resulting in a total of 31 channels.

[1]https://www1.cs.columbia.edu/CAVE/databases/multispectral/
[2]https://repository.kaust.edu.sa/handle/10754/670368

B. Ablation Study

Here we verified the effectiveness of the methodology using the positional encoding (PE) in the residual blocks and the implicit neural representation of PSF and SSF. Four commonly used metrics root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [82], structural similarity index (SSIM) are brought in to evaluate the quality of reconstructed high-resolution HSI. We compared different methods on both CAVE and KAUST datasets.

1) Positional Encoding in Spatial-variant Resblock: We compared three different structures of Γ in resblock: concatenated with/without PE, and concatenated with the spatial coordinates (u, v). Note that the positional encoding vector we used contains only six channels including the two-dimensional coordinates, which are uniformly sampled from the high-dimensional vector $\gamma(p)(L = 64)$.

Table II shows the fusion quality of the three structures. Here, the PSF we used is the 32×32 kernel. It is evident that introducing coordinate information into residual blocks can effectively enhance fusion quality. The PSNRs are improved by nearly 3dB on both two datasets when position encoding is applied.

Specifically, we observed that the PE in the spatial-variant residual block can effectively eliminate the effect of the metamerism in HSI fusion. As shown in Fig.4, the RGB appearances of the real and fake peppers are similar, as well as it is in fake and real faces, but the spectral plots of the two peppers and two faces are different. The PE structure can accurately reconstruct the two spectra, while the residual block without PE can only give results that appear to be a mixture of the two ground truth spectra. This is because the pure resblock without position information is spatial invariant and tends to give average spectral results for points with similar RGB but different spectra.

TABLE II
THE COMPARISON OF FUSION QUALITY OF THE THREE STRUCTURES IN RESBLOCKS.

Datasets	Methods	RMSE↓		PSNR↑		SAM↓		SSIM↑	
		mean	std	mean	std	mean	std	mean	std
CAVE	w/o PE	2.438	0.579	40.568	1.822	6.598	1.714	0.983	0.003
	uv only	1.752	**0.242**	43.334	**1.212**	5.400	**1.375**	0.989	**0.002**
	with PE	**1.696**	0.346	**43.696**	1.768	**5.256**	1.431	**0.990**	0.003
KAUST	w/o PE	1.904	1.097	43.531	4.502	4.242	**1.393**	0.986	0.005
	uv only	1.488	0.620	45.163	**3.076**	3.484	2.143	0.991	0.004
	with PE	**1.300**	**0.564**	**46.440**	3.492	**3.353**	2.101	**0.992**	**0.002**

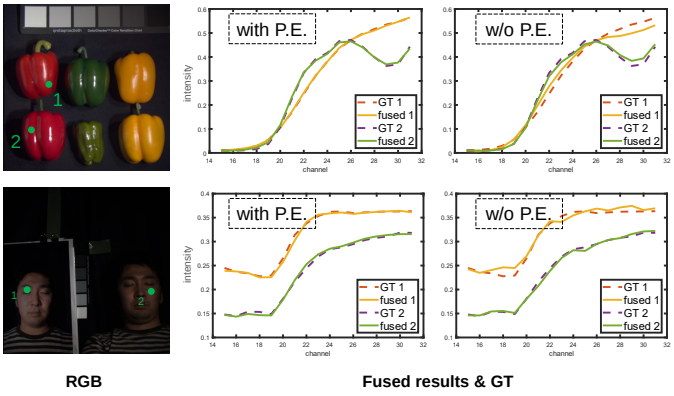


Fig. 4. Metamerism comparison of the results produced by structures with and without positional coding.

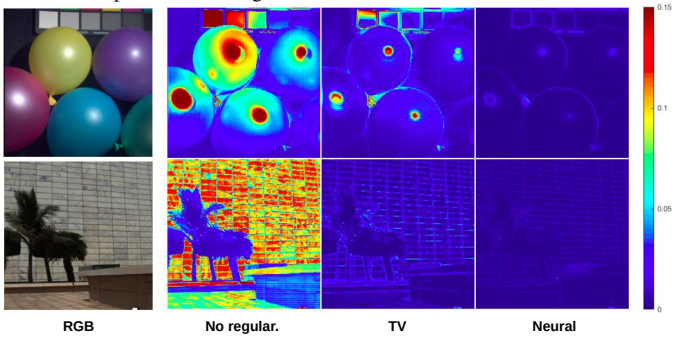


Fig. 5. The effects of total variation (TV) prior and implicit neural representation (Neural) on the errors of the reconstructed high-resolution HSIs on CAVE *balloons* (up) and KAUST *tree_and_wall* (bottom).

2) Implicit Neural Representation of PSF & SSF: First we evaluated four basis functions of positional encoding, which are *sin&cos* (as shown in Eq.7), *rbf*, *sawtooth*, and *dirichlet*. The results are conducted on the CAVE dataset with two kernel sizes of PSF, as shown in Table V. Overall, there is not much difference between the four basis functions, but the *sin&cos* strategy proves to be more stable, therefore we choose it as the used basis function of positional encoding for the subsequent experiments. Then, to further verify the effectiveness of the implicit neural representation of PSF & SSF, we compared it with two other strategies: adding TV regularizers for both PSF and SSF, and representing PSF and SSF as trainable tensors without any regularizer (no regular.). Note that for implicit neural representation, we respectively apply nine and six channels PE tensors for generating SSFs ($L = 4$) and PSFs ($L = 1$). TV is chosen as a comparison method since it is a commonly used regularizer for estimating PSF and SSF [23]. In the TV strategy, we set the fixed weight 0.5×10^{-5} and 1×10^{-4} for the TV terms of PSFs and SSFs respectively.

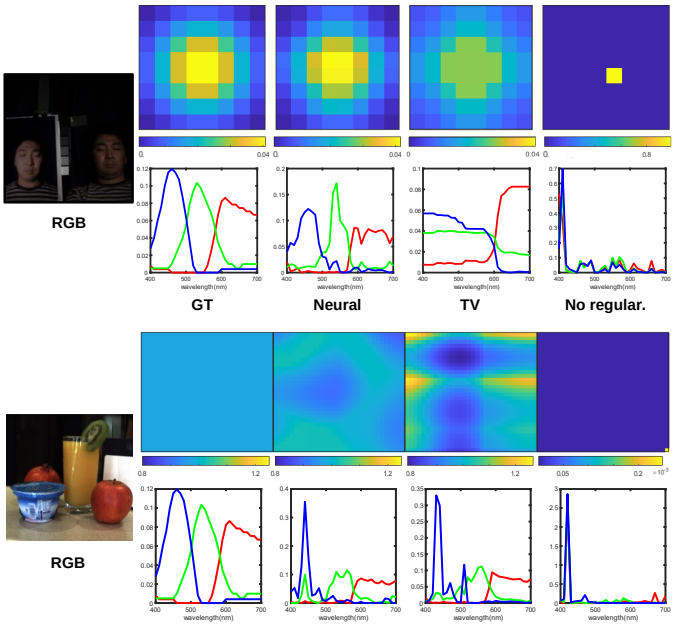


Fig. 6. PSF and SSFs estimated by the three strategies for dealing with degradation model. Here the 8×8 and 32×32 PSFs are used.

We compared the three strategies on the two datasets with the uniform 32×32 PSF. As shown in Table III, our implicit neural representation significantly outperforms others on the four metrics. It proves that the deep prior of the degradation parameters performs better than other handcrafted prior in our network, such as TV. This is because our implicit neural representation has higher optimization freedom than TV, moreover, the implicit constraints in the neural representation are not affected by differences in reprojection errors of different images. Fig.5 shows the error between the GT and the estimated high-resolution HSI by using the three strategies. It is significant that our implicit neural representation method is much more suitable for regularization and more effective to reduce the fusion error.

We showed two examples of estimated PSFs and SSFs of the three strategies with 8×8 and 32×32 PSF kernels in Fig.6. Both our PSF and SSF are close to the ground truth, while other strategies might get trapped into local minima and give inaccurate results.

TABLE III
THE COMPARISON OF THE THREE STRATEGIES FOR DEALING WITH THE DEGRADATION MODEL.

Datasets	Methods	RMSE↓		PSNR↑		SAM↓		SSIM↑	
		mean	std	mean	std	mean	std	mean	std
CAVE	No regular.	15.823	4.019	24.388	2.230	22.577	5.945	0.775	0.048
	TV	9.189	4.171	29.804	4.610	15.147	7.303	0.879	0.075
	Neural	**1.868**	**0.367**	**42.846**	**1.709**	**5.603**	**1.284**	**0.987**	**0.003**
KAUST	No regular.	14.294	8.487	26.772	6.636	13.486	4.079	0.752	0.105
	TV	4.640	3.056	36.500	6.243	6.483	2.665	0.938	0.039
	Neural	**1.555**	**0.500**	**44.622**	**2.584**	**3.887**	**2.450**	**0.989**	**0.002**

TABLE IV
QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT METHODS ON THE CAVE AND KAUST DATASETS WITH TWO TYPES OF PSF. THE BEST METHOD IS HIGHLIGHTED IN BOLD.

Kernel Sizes	Datasets	Metrics	Non-blind					Semi-blind			Blind		
			uSDN	UMAG	NLSTF	LTTR	LRTA	CUCaNet	UAL	HyCoNet	Hysure	DBSR	Ours
8	CAVE	RMSE↓	2.964	1.343	2.041	1.299	2.066	1.607	1.557	1.949	16.73	3.789	**1.284**
		PSNR↑	38.84	45.61	41.97	45.88	41.86	44.08	44.40	42.52	29.77	36.99	**46.04**
		SAM↓	10.74	5.424	4.998	**4.305**	9.705	4.844	4.838	5.166	17.51	7.844	4.493
		SSIM↑	0.964	0.991	0.985	0.992	0.977	0.992	0.991	0.990	0.826	0.976	**0.993**
	KAUST	RMSE↓	2.863	1.104	1.504	1.029	1.858	1.463	2.394	0.794	10.80	6.863	**0.954**
		PSNR↑	39.76	47.74	45.23	48.09	43.68	45.68	41.93	45.47	32.74	34.87	**48.79**
		SAM↓	9.428	3.621	3.993	3.441	5.770	3.890	5.683	3.915	12.22	4.808	**3.427**
		SSIM↑	0.973	0.991	0.987	0.992	0.981	0.991	0.977	0.991	0.885	0.949	**0.993**
32	CAVE	RMSE↓	3.842	2.130	2.131	1.865	3.522	3.058	1.720	3.531	18.57	3.204	**1.696**
		PSNR↑	36.57	41.70	41.65	42.77	37.26	38.52	43.49	37.27	27.11	38.02	**43.70**
		SAM↓	12.60	6.896	6.079	6.059	17.71	9.146	5.414	10.86	19.12	7.076	**5.255**
		SSIM↑	0.948	0.982	0.986	0.988	0.965	0.983	0.989	0.975	0.808	0.977	**0.990**
	KAUST	RMSE↓	5.493	2.025	1.714	1.723	2.948	2.767	2.111	3.405	5.39	2.035	**1.546**
		PSNR↑	33.92	42.76	44.01	43.85	39.59	39.91	43.02	38.79	34.34	41.98	**44.64**
		SAM↓	16.51	5.724	4.794	5.279	9.182	8.686	4.925	8.339	9.27	5.007	**4.808**
		SSIM↑	0.802	0.981	0.986	0.986	0.975	0.977	0.985	0.978	0.940	0.975	**0.988**

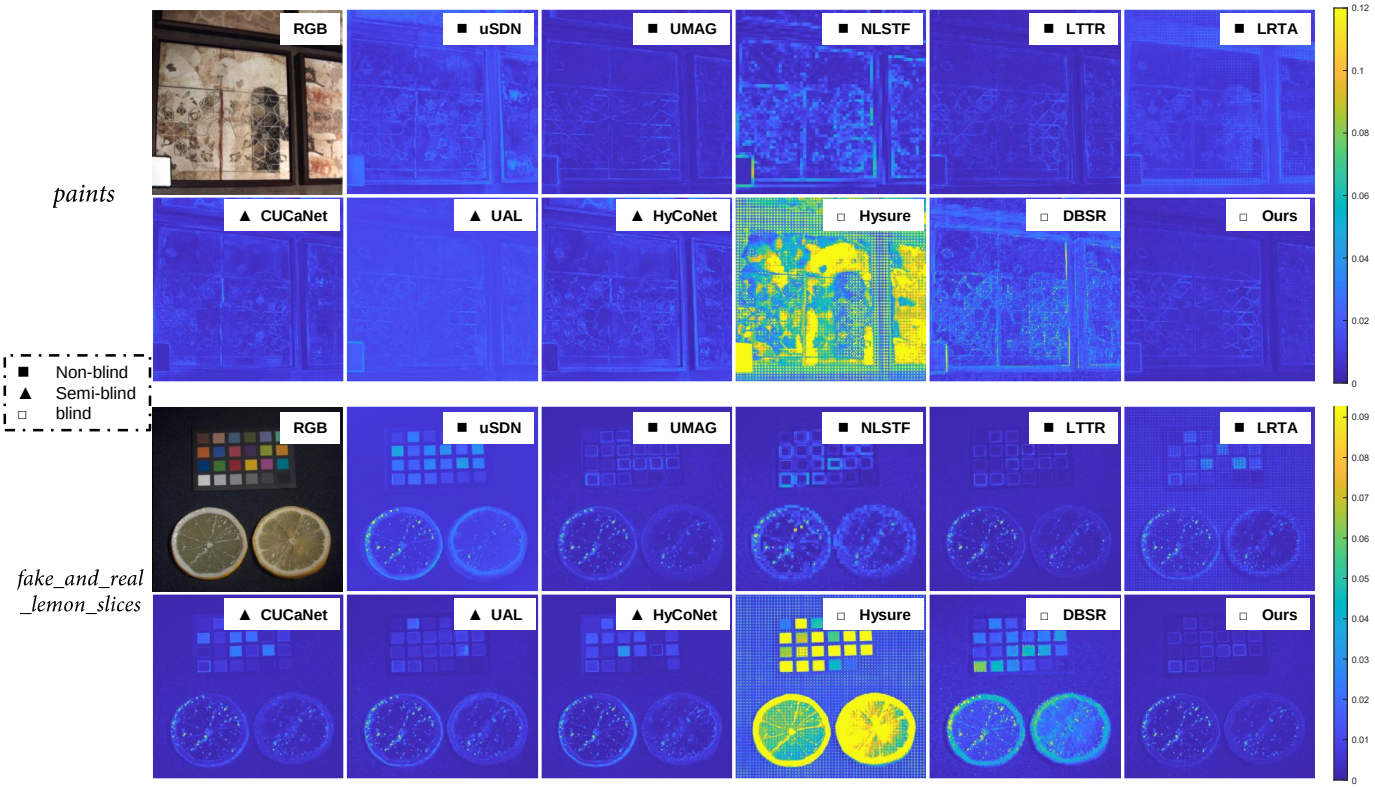
C. Simulation

Then we compared our *BUSI-FusionNet*with several state-of-the-art unsupervised HSI and RGB image fusion methods, including UMAG-Net [47], NLSTF [36][3], uSDN [48][4], LRTA [31][5], LTTR [39][6], CUCaNet [29][7], UAL [30][8], HyCoNet [28][9], Hysure [23][10] and DBSR [24][11] on two synthetic datasets. The first five methods are all non-blind; the CUCaNet, HyCoNet and UAL are used as the semi-blind method, which treats the SSF as a known parameter and estimates the PSF only; while Hysure and DBSR are blind methods like ours. Note that the SSFs in CUCaNet [29] and HyCoNet [28] are not explicit trainable matrices by their default settings, for fair comparisons, we treated their SSFs as known parameters and fixed them to the ground truth, and ran them as semi-blind fusion methods.

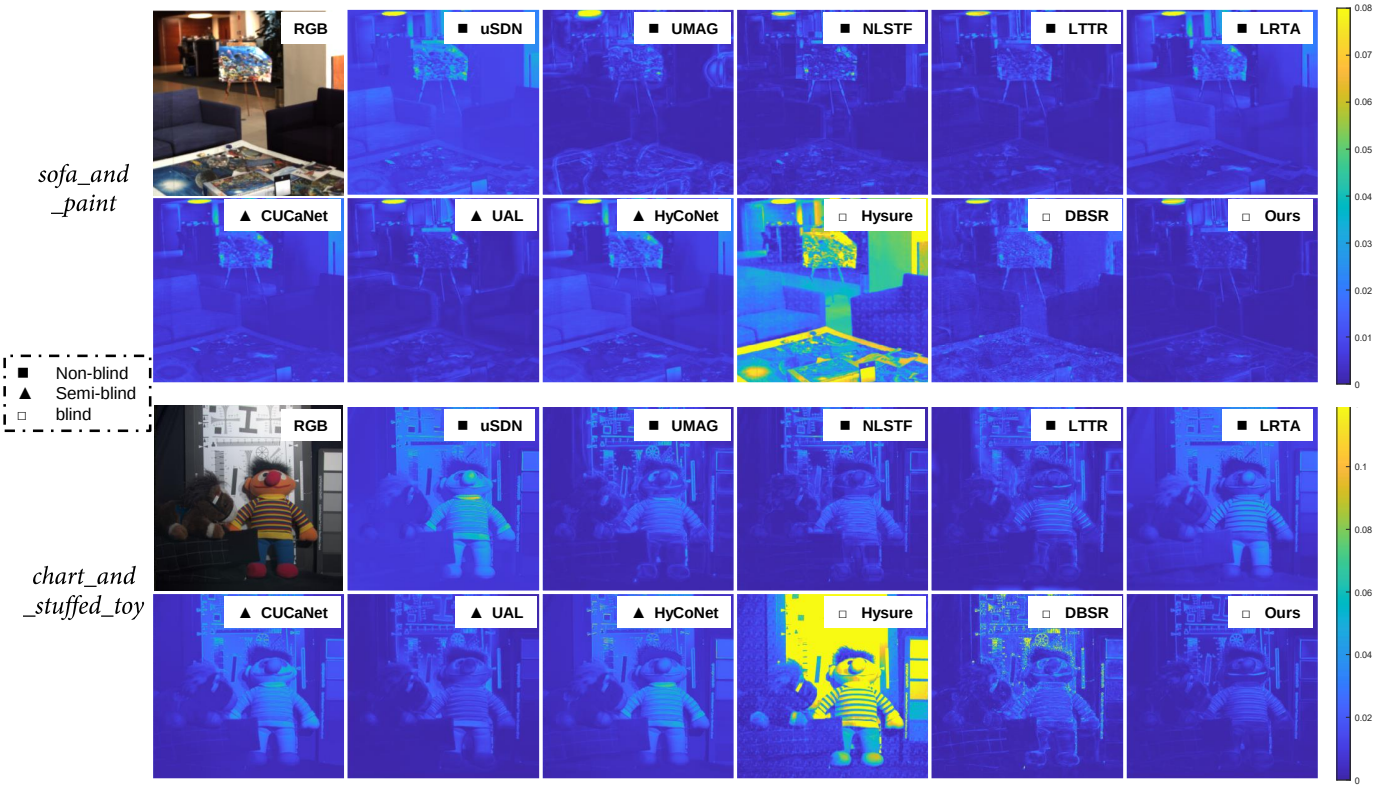
The settings of the comparison methods are conducted by default settings. The PSFs and the SSFs of the blind methods are initialized randomly when they are trainable, except DBSR. We initialized the SSF of DBSR to ground truth since this method is very sensitive to initialization, and easily gets trapped in local minima, making it difficult to generate good SSFs or fused high-resolution HSI.

Table IV shows the comparison results of our method and others. Generally, most methods perform better with 8×8 PSF kernel than with 32×32 PSF, and the numerical results of the metrics on KAUST are better than those on CAVE. Our method outperforms other state-of-the-art unsupervised semi/blind methods, and it is even better than these five non-blind methods on most of the metrics. This illustrates the advantage of our deep image prior compared to other hand-crafted priors, such as sparsity, and nonlocal low-rank. These hand-crafted priors can only express shallow structures and might fail in challenging cases, while our multi-stage residual block can provide sufficient approximation capacity.

[3]https://github.com/renweidian/NLSTF
[4]https://github.com/aicip/uSDN
[5]https://openremotesensing.net/knowledgebase/hyperspectral-restoration-and-fusion-with-multispectral-imagery-by-recasting-low-rank-tensor-approximation/
[6]https://github.com/renweidian/LTTR
[7]https://github.com/danfenghong/ECCV2020_CUCaNet
[8]https://github.com/JiangtaoNie/UAL
[9]https://github.com/saber-zero/HyperFusion
[10]https://github.com/alfaiate/HySure
[11]https://github.com/JiangtaoNie/DBSR



(a) with 8×8 PSF kernel.



(b) with 32×32 PSF kernel.

Fig. 7. The comparison of fusion error of the eleven methods on four scenes with two PSFs.

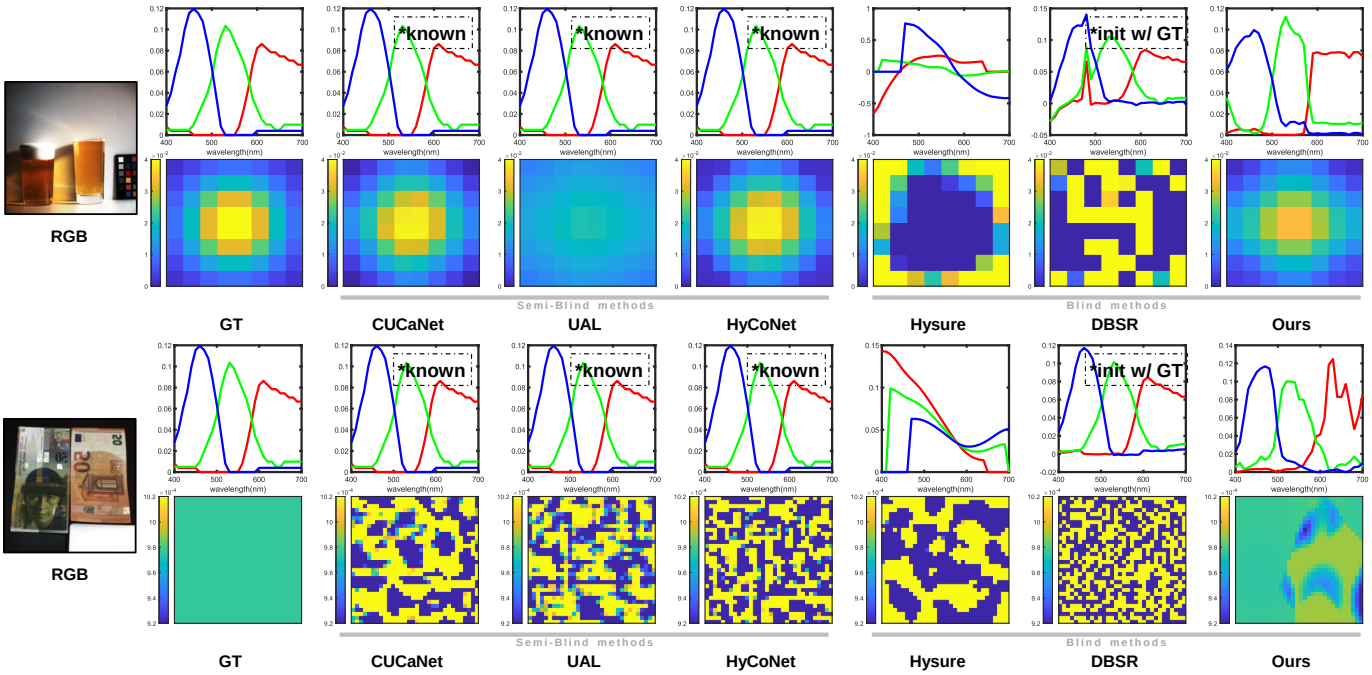


Fig. 8. The comparison of PSF and SSF estimation using the three blind methods with two PSFs (Upper: 8×8 PSF kernel on CAVE dataset; Bottom: 32×32 PSK kernel on KAUST dataset). Note that the SSFs in semi-blind group are as known, and the SSFs of DBSR are initialized to the ground truth SSFs because the DBSR is underperforming with other randomly initialized SSFs.

We visualized the spectral error of the reconstructed high-resolution HSIs of eleven methods in Fig.7. Here the spectral error of each pixel is calculated by using the Euclidean distance between the reconstructed spectrum and the target spectrum. We gave the results on two example scenes, each was reconstructed with the two types of PSFs. Our results have the highest fidelity to the ground truth and perform robustly with the two PSFs.

Fig.8 shows the SSF and PSF estimated by three semi-blind methods and three blind methods (including ours) with the two types of PSFs. Our degradation estimation shows superior performance than the others in the challenging task.

D. Experiments on real data

We also test our *BUSI-FusionNet*on the real image pairs dataset. Before feeding the images into the network, we first roughly match the content of the two images, then crop the RGB images and low-resolution HSI into 1536×1536 and 256×256 patches respectively, lastly we make pixel-to-pixel registration by applying GMA network [78] to estimate high-quality optical flow between the two patches, and warp RGB images to match low-resolution HSI. We used forward and backward optical flow to evaluate the consistency of the flow and guarantee the used image pairs have fewer occlusion regions.

In the fusion step, we set the PSF size to be 6×6, and send the low-resolution HSI patches with the size of 128×128 and the corresponding RGB patches with the size of 768×768 to *BUSI-FusionNet*. Two examples of results are shown in Fig.9.

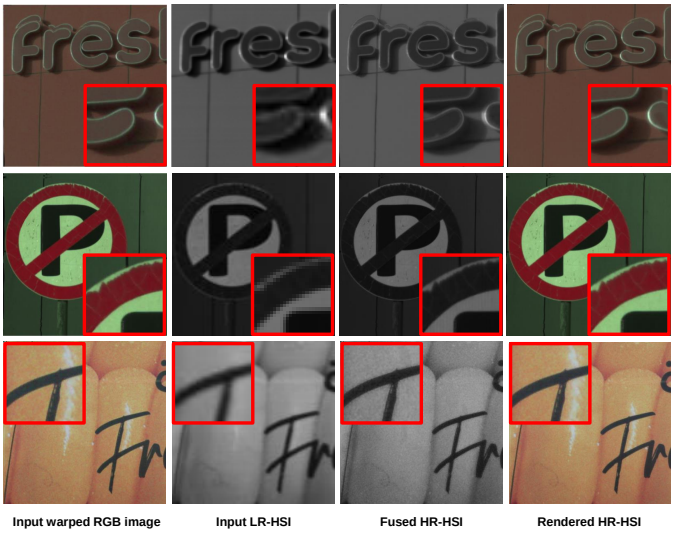


Fig. 9. The results of our proposed *BUSI-FusionNet*on real image dataset. Note that we show the HSIs at the 19^{th} band, and the rendered high-resolution HSIs are obtained by multiplying the fused high-resolution HSIs with estimated SSFs. Details are $3\times$ enlarged in the red box. Note that the details of the highlight regions are well preserved in the fusion result images.

TABLE V
THE COMPARISON OF THE FOUR BASIS FUNCTIONS OF POSITIONAL ENCODING ON THE CAVE DATASET. THE BEST METHOD IS HIGHLIGHTED IN BOLD.

Kernel sizes	Functions	PSNR↑	SAM↓	SSIM↑	RMSE↓
8	dirichlet	44.91	5.278	0.991	1.491
	rbf	43.91	5.314	0.991	1.633
	sawtooth	**46.77**	5.243	0.993	**1.174**
	sin&cos	46.04	**4.493**	**0.993**	1.284
32	dirichlet	43.44	6.251	0.989	1.719
	rbf	43.01	6.417	0.987	1.838
	sawtooth	42.82	6.835	0.987	1.868
	sin&cos	**43.70**	**5.255**	**0.990**	**1.696**

Although ghost artifacts exist in some regions of the warped images due to imperfect optical flow estimation, our method can effectively enhance the spatial and spectral resolution in most regions through fusion. The details such as highlight ad edges are preserved in the recovered high-resolution HSIs.

V. Conclusion

In this paper, we propose the *BUSI-FusionNet* for the task of single image RGB-HSI fusion as well as the PSFs and SRFs estimation without a training set. The network unrolls the optimization of a Sylvester equation, and utilizes a spatially varying denoising network as well as positional encoding to adequately fit the target high-resolution HSIs while resolving metamerism issues. The implicit neural representation of the degradation model shows superior performance compared to handcrafted priors. Several experiments on both the CAVE and KAUST synthetic datasets as well as real images demonstrate the robustness and accuracy of *BUSI-FusionNet*. We construct a large real paired RGB and low-resolution HSI image dataset for evaluation, and the community can use it for future analysis work. However, the proposed method still has some limitations. First, it requires a long training time, and the image registration operation is excluded from our end-to-end network. Therefore, in the future, we will explore a unified framework for joint high-resolution HSI fusion and image registration. We will investigate parameter initialization methods (such as meta-learning) for faster convergence. Also, we would like to try other generative models to handle the challenging unsupervised single blind image fusion task.

References

[1] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, pp. 1552–1565, 2004.

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

[3] Y. Li, A. Majumder, D. Lu, and M. Gopi, "Content-independent multi-spectral display using superimposed projections," in *Computer Graphics Forum*, vol. 34, no. 2. Wiley Online Library, 2015, pp. 337–348.

[4] C. Ma, X. Cao, X. Tong, Q. Dai, and S. Lin, "Acquisition of high spatial and spectral resolution video with a hybrid camera system," *International journal of computer vision*, vol. 110, no. 2, pp. 141–155, 2014.

[5] P.-J. Lapray, X. Wang, J.-B. Thomas, and P. Gouton, "Multispectral filter arrays: Recent advances and practical implementation," *Sensors*, vol. 14, no. 11, pp. 21 626–21 659, 2014.

[6] S.-H. Baek, H. Ikoma, D. S. Jeon, Y. Li, W. Heidrich, G. Wetzstein, and M. H. Kim, "Single-shot hyperspectral-depth imaging with learned diffractive optics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2651–2660.

[7] B. Uzkent, A. Rangnekar, and M. Hoffman, "Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 39–48.

[8] H. Van Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 44–51.

[9] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6440–6461, 2018.

[10] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 2, pp. 29–56, 2017.

[11] E. Maggiori, G. Charpiat, Y. Tarabalka, and P. Alliez, "Recurrent neural networks to correct satellite image classification maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 4962–4971, 2017.

[12] C. Kwan, B. Ayhan, G. Chen, J. Wang, B. Ji, and C.-I. Chang, "A novel approach for spectral unmixing, classification, and concentration estimation of chemical and biological agents," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 2, pp. 409–419, 2006.

[13] P. S. S. Aydav and S. Minz, "Classification of hyperspectral images using self-training and a pseudo validation set," *Remote Sensing Letters*, vol. 9, no. 11, pp. 1109–1117, 2018.

[14] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2014.

[15] X. Fan, H. Rhody, and E. Saber, "A spatial-feature-enhanced mmi algorithm for multimodal airborne image registration," *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 6, pp. 2580–2589, 2010.

[16] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 114–141, 2003.

[17] Y. Zhou, A. Rangarajan, and P. D. Gader, "Nonrigid registration of hyperspectral and color images with vastly different spatial and spectral resolutions for spectral unmixing and pansharpening," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 86–94.

[18] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Information Fusion*, vol. 12, no. 2, pp. 74–84, 2011.

[19] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[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. Simões, J.-Y. Tourneret, M. A. Veganzones, G. Vivone, Q. Wei, and N. Yokoya, "Hyperspectral pansharpening: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 27–46, 2015.

[21] L. Hou and X. Zhang, "Pansharpening image fusion using cross-channel correlation: A framelet-based approach," *Journal of Mathematical Imaging and Vision*, vol. 55, no. 1, pp. 36–49, 2016.

[22] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for p+ xs image fusion," *International Journal of Computer Vision*, vol. 69, no. 1, pp. 43–58, 2006.

[23] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2014.

[24] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2388–2400, 2020.

[25] J. Long and Y. Peng, "Blind fusion of hyperspectral multispectral images based on matrix factorization," *Remote Sensing*, vol. 13, no. 21, p. 4219, 2021.

[26] Z. Liu, Y. Zheng, and X.-H. Han, "Unsupervised multispectral and hyperspectral image fusion with deep spatial and spectral priors," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[27] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geoscience and Remote Sensing Letters*, 2022.

[28] K. Zheng, L. Gao, W. Liao, D. Hong, B. Zhang, X. Cui, and J. Chanussot, "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2487–2502, 2020.

[29] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.

[30] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3073–3082.

[31] N. Liu, L. Li, W. Li, R. Tao, J. E. Fowler, and J. Chanussot, "Hyperspectral restoration and fusion with multispectral imagery via low-rank tensor-approximation," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[32] W. Wan, W. Guo, H. Huang, and J. Liu, "Nonnegative and nonlocal sparse tensor factorization-based hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8384–8394, 2020.

[33] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[34] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, 2015.

[35] X. Li, Y. Zhang, Z. Ge, G. Cao, H. Shi, and P. Fu, "Adaptive nonnegative sparse representation for hyperspectral image super-resolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4267–4283, 2021.

[36] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5344–5353.

[37] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2011.

[38] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by ms/hs fusion net," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1585–1594.

[39] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2672–2683, 2019.

[40] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor cp decomposition for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 348–362, 2019.

[41] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 8028–8042, 2020.

[42] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3586–3594.

[43] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE transactions on image processing*, vol. 29, pp. 116–127, 2019.

[44] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 63–78.

[45] ——, "Bayesian sparse representation for hyperspectral image super resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3631–3640.

[46] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[47] S. Liu, S. Miao, J. Su, B. Li, W. Hu, and Y.-D. Zhang, "Umag-net: A new unsupervised multiattention-guided network for hyperspectral and multispectral image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7373–7385, 2021.

[48] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse dirichlet-net for hyperspectral image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2511–2520.

[49] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, "Unsupervised and unregistered hyperspectral image super-resolution with mutual dirichlet-net," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[50] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, 2021.

[51] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4150–4159.

[52] W. Wang, X. Fu, W. Zeng, L. Sun, R. Zhan, Y. Huang, and X. Ding, "Enhanced deep blind hyperspectral image fusion," *IEEE transactions on neural networks and learning systems*, 2021.

[53] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[54] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 201–214, 2022.

[55] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[56] X.-H. Han, B. Shi, and Y. Zheng, "Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2506–2510.

[57] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 1423–1438, 2020.

[58] T. Zhang, Y. Fu, L. Wang, and H. Huang, "Hyperspectral image reconstruction using deep external and internal learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8559–8568.

[59] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Information Fusion*, vol. 69, pp. 40–51, 2021.

[60] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2091–2106, 2005.

[61] J.-L. Starck, J. Fadili, and F. Murtagh, "The undecimated wavelet decomposition and its reconstruction," *IEEE transactions on image processing*, vol. 16, no. 2, pp. 297–309, 2007.

[62] B. Huang, H. Song, H. Cui, J. Peng, and Z. Xu, "Spatial and spectral image fusion using sparse matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 3, pp. 1693–1704, 2013.

[63] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE transactions on cybernetics*, vol. 50, no. 10, pp. 4469–4480, 2019.

[64] Y. Chang, L. Yan, X.-L. Zhao, H. Fang, Z. Zhang, and S. Zhong, "Weighted low-rank tensor recovery for hyperspectral image restoration," *IEEE transactions on cybernetics*, vol. 50, no. 11, pp. 4558–4572, 2020.

[65] C. Prévost, K. Usevich, P. Comon, and D. Brie, "Hyperspectral super-resolution with coupled tucker approximation: Recoverability and svd-based algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 931–946, 2020.

[66] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[67] ——, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[68] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

[69] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2020.

[70] W. Wei, J. Nie, Y. Li, L. Zhang, and Y. Zhang, "Deep recursive network for hyperspectral image super-resolution," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1233–1244, 2020.

[71] J. Yang, L. Xiao, Y.-Q. Zhao, and J. C.-W. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

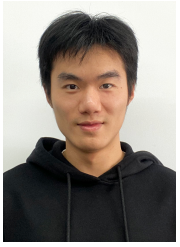
[72] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[73] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized rgb guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 661–11 670.

[74] Y. Li, C. Wang, and J. Zhao, "Locally linear embedded sparse coding for spectral reconstruction from rgb images," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 363–367, 2017.

[75] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE transactions*

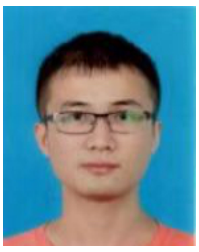
on pattern analysis and machine intelligence, vol. 41, no. 10, pp. 2305–2318, 2018.
[76] Y. Li, M. Qi, R. Gulve, M. Wei, R. Genov, K. N. Kutulakos, and W. Heidrich, "End-to-end video compressive sensing using anderson-accelerated unrolled networks," in *2020 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2020, pp. 1–12.
[77] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.
[78] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," *arXiv preprint arXiv:2104.02409*, 2021.
[79] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conf. on Computer Vision (ECCV)*, ser. Part IV, LNCS 7577, A. Fitzgibbon et al. (Eds.), Ed. Springer-Verlag, Oct. 2012, pp. 611–625.
[80] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum," *IEEE transactions on image processing*, vol. 19, no. 9, pp. 2241–2253, 2010.
[81] Y. Li, Q. Fu, and W. Heidrich, "Multispectral illumination estimation using deep unrolling network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2672–2681.
[82] J. Boardman, "Spectral angle mapping: a rapid measure of spectral similarity," *AVIRIS. Delivered by Ingenta*, 1993.



Jiabao Li Jiabao Li received the B.S. degree in Electronic Information Science and Technology from Wenzhou University, Wenzhou, China, in 2020. He is currently working toward the M.S. degree in Computer Technology at Ningbo University, Ningbo, China. His research interests include computational imaging and deep learning.

Yuqi Li Yuqi Li is an associate professor at Ningbo University. He received the B.Sc. degree and the Ph.D. degree in computer science and technology from Zhejiang University in 2010 and 2016. His recent interest is in computational imaging and computational display, focusing on high-color-fidelity display, image reconstruction, and deep learning.

Chong Wang Chong Wang (S'12) received his B. Eng degree from Zhejiang University of Technology in 2007, M. Eng degree from University of Science and Technology of China in 2010, and his Ph.D. degree from The University of Hong Kong in 2014. His main research interests are in depth camera assisted systems, gesture recognition, image and video restoration, image based rendering and parallel computing.



Xulun Ye Xulun Ye received an M.Sc. and PhD degree from Ningbo University, China, in 2016 and 2019, where he is currently a lecturer. His research interests include Bayesian learning, nonparametric clustering and convex analysis.

Wolfgang Heidrich Wolfgang Heidrich is a Professor of Computer Science and the Director of the Visual Computing Center at King Abdullah University of Science and Technology. He received his PhD in Computer Science from the University of Erlangen in 1999, and then worked as a Research Associate in the Computer Graphics Group of the Max Planck Institute for Computer Science in Saarbrucken, Germany, before joining the faculty of the University of British Columbia in 2000, initially as an Assistant, then Associate and Full Professor, and finally Dolby Research Chair. In 2014, he joined King Abdullah University of Science and Technology while continuing to affiliated with University of British Columbia until 2018. His research interests lie at the intersection of computer graphics, computer vision, imaging, and optics. In particular, he has worked on computational imaging and displays, high dynamic range imaging and display, image-based modeling, measuring, and rendering, geometry acquisition, GPUbased rendering, and global illumination. He has written well over 200 refereed publications on these subjects and has served on numerous program committees. His work on High Dynamic Range Displays served as the basis for the technology behind Brightside Technologies, which was acquired by Dolby in 2007. In 2016, he was the papers chair for both SIGGRAPH ASIA and ICCP. He is the recipient of a 2014 Humboldt Research Award.