“Multispectral and Hyperspectral Image Fusion by MS/HS Fusion Net”: Supplementary Material

Anonymous CVPR submission
Paper ID 60

Abstract

*In this supplementary material, we provide the proofs to **Theorems 1** and **Corollary 1** presented in the main text. We then present more details on the implementation and the parameter settings of our method in our experiments. We also provide more analyses about the proposed MS/HS fusion model. Finally, we show more experimental results for model verification and more comprehensive performance demonstrations.*

1. Proofs to Theorem 1 and Corollary 1

We now prove **Theorem 1** in the main text.

Theorem 1. *For any $\boldsymbol{X} \in \mathbb{R}^{HW\times S}$ and $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{HW\times s}$, if* $\text{rank}(\boldsymbol{X}) = r > s$ *and* $\text{rank}(\tilde{\boldsymbol{Y}}) = s$*, then the following two statements are equivalent to each other:*
(a) There exists an $\boldsymbol{R} \in \mathbb{R}^{S\times s}$, subject to,

$$\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{R}. \tag{1}$$

(b) There exist $\boldsymbol{A} \in \mathbb{R}^{s\times S}$, $\boldsymbol{B} \in \mathbb{R}^{(r-s)\times S}$ and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{HW\times(r-s)}$, subject to,

$$\boldsymbol{X} = \tilde{\boldsymbol{Y}}\boldsymbol{A} + \hat{\boldsymbol{Y}}\boldsymbol{B}. \tag{2}$$

Proof. 1). We first prove that when (b) is satisfied, (a) can be deduced.

Let $\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix}$, and then we have

$$\boldsymbol{X} = \left[\tilde{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}\right]\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix} = \left[\tilde{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}\right]\boldsymbol{Q}. \tag{3}$$

By (3), we can obtain that $\text{rank}(\boldsymbol{Q}) \geq \text{rank}(\boldsymbol{X}) = r$. Moreover, since $\boldsymbol{Q} \in \mathbb{R}^{r\times S}$, we have $\text{rank}(\boldsymbol{Q}) \leq r$. Thus, we have $\text{rank}(\boldsymbol{Q}) = r$.

We can now prove that $\boldsymbol{Q}\boldsymbol{Q}^T$ is invertible, since $\text{rank}\left(\boldsymbol{Q}\boldsymbol{Q}^T\right) = \text{rank}(\boldsymbol{Q}) = r$. Then, by (2), we have

$$\left[\tilde{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}\right] = \left[\tilde{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}\right]\boldsymbol{Q}\boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T\right)^{-1} = \boldsymbol{X}\boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T\right)^{-1}. \tag{4}$$

Set $\boldsymbol{R} \in \mathbb{R}^{S\times s}$ to be the first s columns of $\boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T\right)^{-1}$, and then it is easy to find that $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{R}$, i.e., $\boldsymbol{R}$ satisfies (a).

2). We then prove that when (a) is satisfied, (b) can be deduced.

Since $\text{rank}(\boldsymbol{X}) = r$, there exist $\boldsymbol{W} \in \mathbb{R}^{HW\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{S\times r}$, s.t.,

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{V}^T. \tag{5}$$

Let $\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{R}$, and then $\boldsymbol{U} \in \mathbb{R}^{r\times s}$, and its singular value decomposition (SVD) is

$$\boldsymbol{U} = \bar{\boldsymbol{U}}\begin{bmatrix}\boldsymbol{\Sigma}\\ \boldsymbol{0}\end{bmatrix}\bar{\boldsymbol{V}}^T, \tag{6}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{s\times s}$ is a diagonal matrix with non-zero diagonal elements, $\boldsymbol{0}$ is an $(r-s)\times(r-s)$ zero matrix, $\bar{\boldsymbol{U}} \in \mathbb{R}^{r\times r}$ and $\bar{\boldsymbol{V}} \in \mathbb{R}^{s\times s}$ are orthogonal matrices.

Denote $\hat{\boldsymbol{U}}$ as the last $r-s$ columns in $\bar{\boldsymbol{U}}$. Then, we have

$$\left[\boldsymbol{U}, \hat{\boldsymbol{U}}\right] = \bar{\boldsymbol{U}}\begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\bar{\boldsymbol{V}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}^T, \tag{7}$$

where $\boldsymbol{I}$ is an $(r-s)\times(r-s)$ identity matrix. It is easy to find that (7) is the SVD of $\left[\boldsymbol{U}, \hat{\boldsymbol{U}}\right]$, and all the singular values are non-zeroes. Therefore, $\left[\boldsymbol{U}, \hat{\boldsymbol{U}}\right]$ is invertible.

Let $\boldsymbol{Q} = \left[\boldsymbol{U}, \hat{\boldsymbol{U}}\right]^{-1}\boldsymbol{V}^T$, then $\boldsymbol{Q} \in \mathbb{R}^{r\times S}$, and we can obtain

$$\begin{aligned}\boldsymbol{X} &= \boldsymbol{W}\boldsymbol{V}^T\\ &= \boldsymbol{W}\left[\boldsymbol{U}, \hat{\boldsymbol{U}}\right]\left[\boldsymbol{U}, \hat{\boldsymbol{U}}\right]^{-1}\boldsymbol{V}^T\\ &= \left[\boldsymbol{W}\boldsymbol{V}^T\boldsymbol{R}, \boldsymbol{W}\hat{\boldsymbol{U}}\right]\boldsymbol{Q}\\ &= \left[\tilde{\boldsymbol{Y}}, \boldsymbol{W}\hat{\boldsymbol{U}}\right]\boldsymbol{Q}.\end{aligned} \tag{8}$$

Let $\hat{\boldsymbol{Y}} = \boldsymbol{W}\hat{\boldsymbol{U}} \in \mathbb{R}^{HW\times(r-s)}$, $\boldsymbol{A} \in \mathbb{R}^{s\times S}$ be the first s rows in $\boldsymbol{Q}$ and $\boldsymbol{B}$ be the last $r-s$ rows in $\boldsymbol{Q}$, and then (8) can be rewriten as $\boldsymbol{X} = \tilde{\boldsymbol{Y}}\boldsymbol{A} + \hat{\boldsymbol{Y}}\boldsymbol{B}$, i.e., (b) is satisfied. □

Then we prove **Corollary** 1 in the main text.

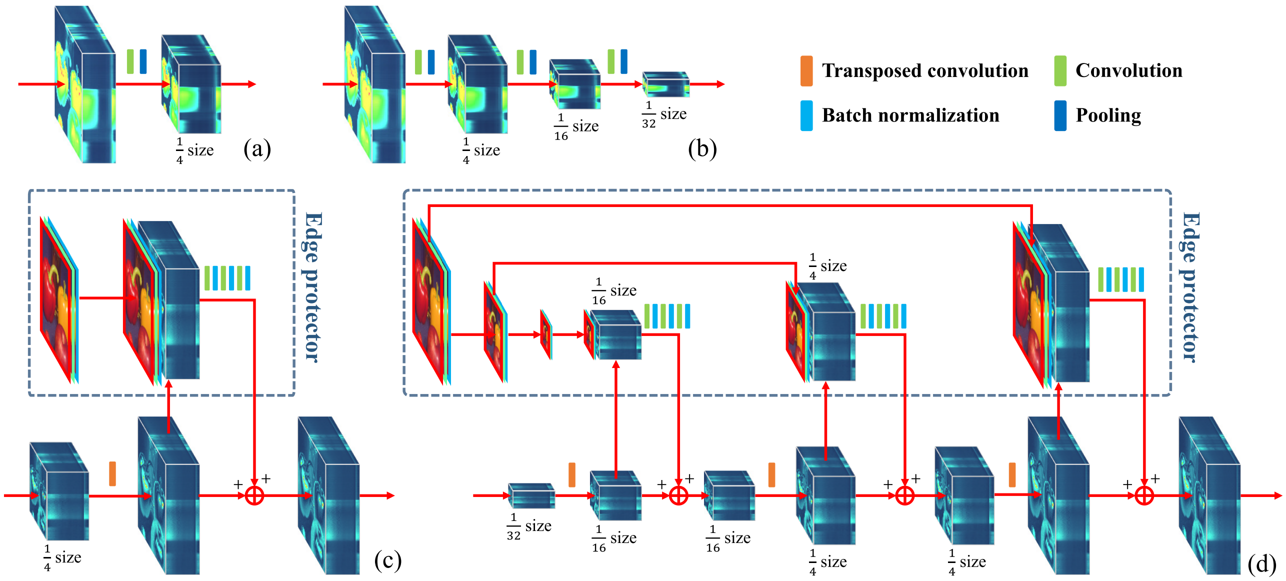


Figure 1. (a) An example of downsampling network with a factor of 4. (b) An example of downsampling network with a factor of 32. (c) An example of upsampling network with a factor of 32. (d) An example of upsampling network with a factor of 32.

Corollary 1. *For any* $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{HW\times s}$, $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{hw\times S}$, $\boldsymbol{C} \in \mathbb{R}^{hw\times HW}$, *if* $\text{rank}(\tilde{\boldsymbol{Y}}) = s$ *and* $\text{rank}(\tilde{\boldsymbol{Z}}) = r > s$, *then the following two statements are equivalent to each other:*
(a) There exist $\boldsymbol{X} \in \mathbb{R}^{HW\times S}$ *and* $\boldsymbol{R} \in \mathbb{R}^{S\times s}$, *subject to,*

$$\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{R}, \;\; \tilde{\boldsymbol{Z}} = \boldsymbol{C}\boldsymbol{X}, \;\; \text{rank}(\boldsymbol{X}) = r. \tag{9}$$

(b) There exist $\boldsymbol{A} \in \mathbb{R}^{s\times S}$, $r > s$, $\boldsymbol{B} \in \mathbb{R}^{(r-s)\times S}$ *and* $\hat{\boldsymbol{Y}} \in \mathbb{R}^{HW\times(r-s)}$, *subject to,*

$$\tilde{\boldsymbol{Z}} = \boldsymbol{C}\left(\tilde{\boldsymbol{Y}}\boldsymbol{A} + \hat{\boldsymbol{Y}}\boldsymbol{B}\right). \tag{10}$$

Proof. 1). We first prove that when (a) is satisfied, (b) can be deduced.

By **Theorem** 1, we know that there exist $\boldsymbol{A} \in \mathbb{R}^{s\times S}$, $\boldsymbol{B} \in \mathbb{R}^{(r-s)\times S}$ and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{(HW\times(r-s))}$, s.t., (2) is satisfied. By combining (2) and $\tilde{\boldsymbol{Z}} = \boldsymbol{C}\boldsymbol{X}$, we can obtain (10), i.e., (b) is satisfied.

2). We then prove that when (b) is satisfied, (a) can be deduced.

Let $\boldsymbol{X} = \tilde{\boldsymbol{Y}}\boldsymbol{A} + \hat{\boldsymbol{Y}}\boldsymbol{B}$, and then we have $\tilde{\boldsymbol{Z}} = \boldsymbol{C}\boldsymbol{X}$ and $\text{rank}(\boldsymbol{X}) \leq r$. Moreover, since $\tilde{\boldsymbol{Z}} = \boldsymbol{C}\boldsymbol{X}$, we can obtain that $\text{rank}(\boldsymbol{X}) \geq \text{rank}(\boldsymbol{Z}) = r$. Therefore, $\text{rank}(\boldsymbol{X}) = r$. In addition, by **Theorem** 1, there exists an $\boldsymbol{R} \in \mathbb{R}^{S\times s}$, s.t., $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{R}$. Therefore, (a) is satisfied by $\boldsymbol{X} = \tilde{\boldsymbol{Y}}\boldsymbol{A} + \hat{\boldsymbol{Y}}\boldsymbol{B}$. □

2. More details of the network design

In this section, we provide more details on the network design of the $\text{downSample}_{\theta_d^{(k)}}(\cdot)$, $\text{upSample}_{\theta_u^{(k)}}(\cdot)$, $\text{proxNet}_{\theta_p^{(k)}}(\cdot)$ and $\text{resNet}_{\theta_r}(\cdot)$.

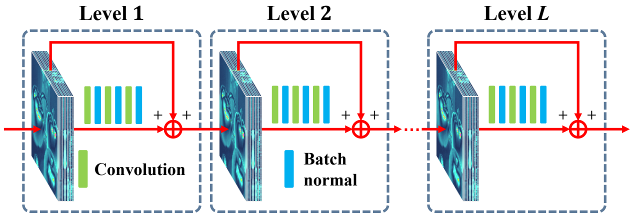


Figure 2. An illustration of exploited ResNet.

Downsampling and upsamling networks. For relatively small scale of factor downsampling and upsampling cases, we simply use a 2D channel-wise convolution operator and an average pooling operator to perform $\text{downSample}_{\theta_d^{(k)}}(\cdot)$ and a 2D transposed convolution to perform $\text{upSample}_{\theta_u^{(k)}}(\cdot)$. One can see Fig. 1 (a) and (b) for easy understanding.

For relatively large-scale of factor downsampling and upsampling cases, such as those with a factor of 32, the simple upsampling result with a 2D transposed convolution can be very blur, which is caused by the fact that the spatial detail information of the image is badly damaged in the large factor downsampling. To address this problem, we use several 4 times spatial downsampling/upsampling and 2 times spatial downsampling/upsampling to approach the large times spatial downsampling/upsampling. Moreover, in $\text{upSample}_{\theta_u^{(k)}}(\cdot)$, we use a 3-level convolution network to restore the spatical details. Specifically, we downsample the HrMS image into a proper size and stuck it with the upsampling result of each stage, and use it as input of the 3-level convolution network. Fig. 1 (c) and (d) show an example of downsampling and upsampling with a factor of 32

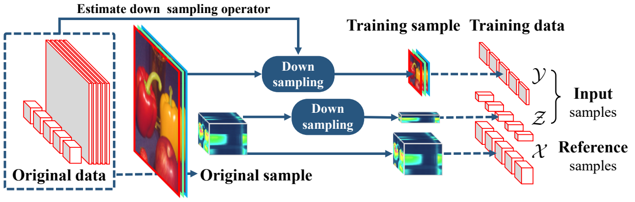


Figure 3. An illustration of how to create the training data when the HrHS images are unavailable.

for easy understanding.

Proximal network and ResNet. We adopt the deep residual network (ResNet) [4] to build the proximal network $\text{proxNet}_{\theta_p^{(k)}}(\cdot)$ and adjusting network $\text{resNet}_{\theta_r}(\cdot)$. Fig. 2 is an illustration of the exploited ResNet.

3. Details of the training data generation

For simulation data and real data where ground truth HrHS images are available, we can easily use the paired training data $\{(\mathcal{Y}_n, \mathcal{Z}_n), \mathcal{X}_n\}_{n=1}^N$ to learn the parameters in the proposed MHF-net.

Unfortunately, in real world, HrHS images $\mathcal{X}_n$s are sometimes unavailable. In this case, we use the method proposed in [10] to address this problem, where the Wald protocol [21] is used to create the training data. Fig. 3 is an illustration of how to create the training data. We downsample both HrMS images and LrHS images, so that the original LrHS images can be taken as references for the downsampled data.

In order to match the sensor properties, we first estimate the spatial downsampling operator $\boldsymbol{C}$ with the observed HrMS images and LrHS images. Specifically, we represent the downsampling operator as

$$\boldsymbol{C}(\cdot) = D(\boldsymbol{\phi} \otimes (\cdot)), \tag{11}$$

where $D(\cdot)$ is a fixed downsampling operator, and $\boldsymbol{\phi} \in \mathbb{R}^{p \times p}$ is a blur kernel matrix. We then estimate $\boldsymbol{\phi}$ by solving the following problem:

$$\begin{aligned} &\min_{\boldsymbol{R},\boldsymbol{\phi}} \sum_n \|\mathcal{Z}_n \times_3 \boldsymbol{R}^T - D(\boldsymbol{\phi} \otimes \mathcal{Y}_n)\|_F^2, \\ &\text{s.t.}, \sum_{i,j} \boldsymbol{\phi}_{ij} = 1, \end{aligned} \tag{12}$$

where $\mathcal{Z}_n$ and $\mathcal{Y}_n$ are the n^{th} observed LrHS and HrMS image samples, and $\boldsymbol{R}$ is the to-be-estimated spectral response of the multispectral sensor. We solve (12) by alternately updating $\boldsymbol{R}$ and $\boldsymbol{\phi}$.

With $\boldsymbol{\phi}$ fixed, $\boldsymbol{R}$ can be updated by solving the following sub-problem:

$$\min_{\boldsymbol{R}} \sum_n \|\boldsymbol{Z}_n \boldsymbol{R} - \text{unfold}_3(D(\boldsymbol{\phi} \otimes \mathcal{Y}_n))\|_F^2, \tag{13}$$

where $\text{unfold}_3(\cdot)$ is the unfolding operator along the third mode and $\boldsymbol{Z}_n = \text{unfold}_3(\mathcal{Z}_n)$. This is a simple least square optimization with closed form solution:

$$\boldsymbol{R}^+ = \left(\sum_n \boldsymbol{Z}_n^T \boldsymbol{Z}_n\right)^{-1} \sum_n \left(\boldsymbol{Z}_n^T \text{unfold}_3\left(D\left(\boldsymbol{\phi} \otimes \mathcal{Y}_n\right)\right)\right). \tag{14}$$

With $\boldsymbol{R}$ fixed, let $\boldsymbol{v} = \text{vec}(\boldsymbol{\phi})$, and then $\boldsymbol{\phi}$ can be updated by solving the following sub-problem with respect to $\boldsymbol{v}$:

$$\begin{aligned} &\min_{\boldsymbol{u}} \|\boldsymbol{w} - \boldsymbol{U}\boldsymbol{v}\|_2^2, \\ &\text{s.t.}, \boldsymbol{1}^T \boldsymbol{v} = 1, \end{aligned} \tag{15}$$

where $\text{vec}(\cdot)$ is the vectorization operator, $\boldsymbol{w}$ is the vector of all elements in $\left\{\mathcal{Z}_n \times_3 \boldsymbol{R}^T\right\}_{n=1}^N$, which is defined by $\boldsymbol{w} = [\boldsymbol{w}_1; \boldsymbol{w}_2; \ldots; \boldsymbol{w}_N]$ with $\boldsymbol{w}_n = \text{vec}\left(\mathcal{Z}_n \times_3 \boldsymbol{R}^T\right)$, and $\boldsymbol{U}$ is the matrix of all patches in $\{\mathcal{Y}_n\}_{n=1}^N$ corresponding to the downsampling operator $D(\cdot)$. To solve problem (15), we first prove the following lemma:

Lemma 1. *The closed-form solution of (15) is:*

$$\boldsymbol{v}^* = \left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1} \left(\boldsymbol{U}^T\boldsymbol{w} - \frac{\boldsymbol{1}^T\left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{w} - 1}{\boldsymbol{1}^T\left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{1}}\boldsymbol{1}\right). \tag{16}$$

Proof. Let $\lambda^* = \frac{2(\boldsymbol{1}^T(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{w} - 1)}{\boldsymbol{1}^T(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{1}}$, and then it is easy to find that $\boldsymbol{v}^*$ and λ^* satisfy the Karush-Kuhn-Tucker (KKT) conditions for convex problem (15), that is:

$$\begin{aligned} &\boldsymbol{1}^T\boldsymbol{v}^* = 1 \\ &\nabla(\|\boldsymbol{w} - \boldsymbol{U}\boldsymbol{v}^*\|_2^2) + \lambda^*\nabla(\boldsymbol{1}^T\boldsymbol{v}^*) \\ &= 2\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{v}^* - 2\boldsymbol{U}^T\boldsymbol{w} + \lambda^*\boldsymbol{1} \\ &= 0. \end{aligned} \tag{17}$$

Therefore, $\boldsymbol{v}^*$ and λ^* are primal and dual optimal, with zero duality gap [3]. □

We can thus update $\boldsymbol{\phi}$ by

$$\boldsymbol{\phi}^+ = \text{fold}_3(\boldsymbol{v}^*). \tag{18}$$

In summary, by alternately performing (14) and (18), we can solve the problem (12), and obtain the downsampling operator. Then we can use the method in Fig. 3 to generate the training data when HrHS images are unavailable.

4. Implementation details in network training

In our method, we implement and train our network using TensorFlow[1] framework. We use Adam optimizer to

[1] https://tensorflow.google.cn/

train the network for 50000 iterations with a batch size of 10 and a learning rate of 0.0001.

We easily set the trade-off parameters α and β in the loss function as 0.1 and 0.01, respectively, and set the rank parameter r as $\min\{15, S\}$, where S is the total band number of the HrHS image. We initialize the parameter $\boldsymbol{A}$ by solving

$$\boldsymbol{A} = (\bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}})^{-1}\bar{\boldsymbol{Y}}^T\bar{\boldsymbol{X}}, \tag{19}$$

where $\bar{\boldsymbol{Y}}$ and $\bar{\boldsymbol{X}}$ are matrices obtained by stacking all the HrMS and HrHS images in the training data along the spatial dimension. It should be noted that (19) is a closed form solution of following problem

$$\min_{\boldsymbol{A}} \|\bar{\boldsymbol{Y}}\boldsymbol{A} - \bar{\boldsymbol{X}}\|_F^2. \tag{20}$$

Besides, we initialize the filters in the donwsampling net $\text{downSample}_{\theta_d^{(k)}}(\cdot)$ and upsampling net $\text{upSample}_{\theta_u^{(k)}}(\cdot)$ with $p \times p$ matrices whose elements are all $\frac{1}{p^2}$, where p is the size of the filter. We initialize the other parameters involved in MHF-net with zero-mean Gaussion distribution with standard deviation 0.1. Our network can perform consistently well and outperform all other competing methods throughout all our experiments under such simple settings.

5. More experimental results

In this section, we provide more experimental results and detail implementations on the three data-set exploited in the main text.

Comparison methods. The comparison methods include: FUSE [16][2], ICCV15 [7][3], GLP-HS [11][4], SFIM-HS [8][4], GSA [1][4], CNMF [19][5], M-FUSE [15][6] and SASFM [5][7], representing the state-of-art traditional methods. Moreover, to better verify the efficiency of the proposed network structure, we implement a network for MS/HS fusion for competition, which only uses the ResNet in the proposed network without using other structures in MHF-net. This method is simply denoted as 'ResNet'. In this method, we set the input as $[\mathcal{Y}, \mathcal{Z}_{up}]$, where $\mathcal{Z}_{up}$ is obtained by interpolating the LrHS image $\mathcal{Z}$ (using a bicubic filter) to the dimension of $\mathcal{Y}$ as [9] did. We set the level number of ResNet to be 30.

Evaluation measures. Five quantitative picture quality indices (PQI) are employed for performance evaluation, including peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [20], erreur relative globale adimensionnelle de synthèse (ERGAS [12]), structure similarity (SSIM [14]), feature similarity (FSIM [22]). SAM calculates the average angle between spectrum vectors of the target MSI and the reference one across all spatial positions and ERGAS measures fidelity of the restored image based on the weighted sum of MSE in each band. PSNR, SSIM and FSIM are conventional PQIs in image processing and computer vision. They evaluate the similarity between the target image and the reference image based on MSE and structural consistency, perceptual consistency, respectively. The smaller ERGAS and SAM are, the better the fusion result is, while the larger PSNR, SSIM and FSIM are, the closer the fusion result is to the reference one.

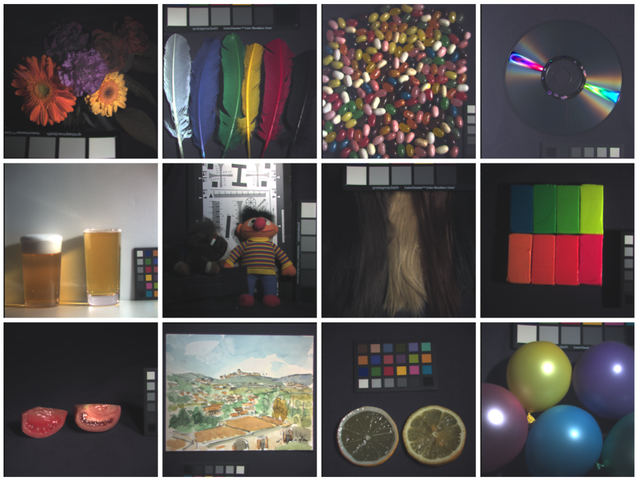


Figure 4. An illustration of the simulated RGB images of the 12 testing samples in CAVE data.

5.1. More results on CAVE data

We first verify the efficiency of the proposed MHF-net on the CAVE Multispectral Image Database [17][8].

The database consists of 32 scenes with spatial size of 512×512, including full spectral resolution reflectance data from 400nm to 700nm at 10nm steps (31 bands in total). We generate the HrMS image (RGB image) by integrating all the ground truth HrHS images with the same simulated spectral response $\boldsymbol{R}$, and generate the LrHS images via downsampling the groundtruth with a factor of 32 implemented by averaging over 32×32 pixel blocks as [2, 6].

To prepare samples for training MHF-net, we randomly select 20 HS images from CAVE database and extract 96×96 overlapped patches from them as reference HrHS images for training. Then the utilized HrHS, HrMS and LrHS images are of size $96 \times 96 \times 31$, $96 \times 96 \times 3$ and $3 \times 3 \times 31$, respectively. The remaining 12 HS images of the database, shown as Fig. 4, are used for validation, where the original images are treated as ground truth HrHS images, and the HrMS and LrHS image are generated in the

[2] http://wei.perso.enseeiht.fr/publications.html

[3] https://github.com/lanha/SupResPALM

[4] http://openremotesensing.net/knowledgebase/hyperspectral-and-multispectral-data-fusion/

[5] http://naotoyokoya.com/Download.html

[6] https://github.com/qw245/BlindFuse

[7] The code was implemented for comparison

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

Table 1. Performance comparison of the competing methods on 12 testing samples in CAVE data set with respect to 5 PQIs.

Data #	1	2	3	4	5	6	7	8	9	10	11	12	Average
PSNR results of the 12 testing data (Ideal value: $+\infty$)													
FUSE	35.83	30.14	28.76	25.17	29.56	19.13	29.91	37.64	33.69	32.45	34.31	34.79	30.95
ICCV15	36.19	29.44	29.23	26.48	29.59	30.07	41.04	36.96	36.06	32.00	34.71	33.48	32.94
GLP-HS	35.22	30.38	29.44	26.56	31.41	30.80	38.84	38.17	35.95	32.55	34.57	32.98	33.07
SFIM-HS	33.54	29.34	25.34	26.83	33.85	30.59	36.61	38.00	34.43	29.30	31.95	32.57	31.86
GSA	36.32	30.95	30.23	25.87	34.43	31.91	39.55	37.69	35.94	32.72	35.62	34.14	33.78
CNMF	36.43	32.13	28.30	26.12	32.27	32.47	40.36	38.68	36.66	32.13	35.12	32.38	33.59
M-FUSE	35.11	29.62	25.69	26.98	34.12	31.43	34.55	35.92	32.90	31.61	31.82	35.53	32.11
SASFM	29.61	24.73	23.78	12.92	28.27	20.58	30.33	29.64	34.16	27.90	31.31	25.89	26.59
PNN	35.07	31.03	30.16	26.48	30.99	31.47	36.34	35.40	35.08	29.11	34.71	33.26	32.43
3D-CNN	36.94	32.21	32.52	26.47	34.71	33.17	42.49	35.28	38.84	33.80	38.24	33.23	34.82
ResNet	34.42	31.44	30.26	25.16	26.81	30.14	38.86	36.08	37.53	27.17	38.23	30.86	32.25
MHF-net	**38.59**	**34.98**	**33.76**	**28.43**	**36.64**	**34.57**	**43.70**	**41.68**	**41.02**	**34.77**	**42.03**	**36.57**	**37.23**
SAM results of the 12 testing data (Ideal value: 0)													
FUSE	18.24	12.47	16.01	10.68	4.49	24.83	14.56	15.10	12.43	8.00	13.11	6.88	13.07
ICCV15	14.58	13.29	11.68	12.85	2.47	13.88	7.44	13.66	8.29	5.76	10.43	7.85	10.18
GLP-HS	17.89	16.27	13.40	11.99	2.50	14.07	10.28	15.44	11.21	6.76	12.79	6.38	11.58
SFIM-HS	**9.12**	8.63	14.43	**8.12**	**1.77**	8.07	6.74	**8.51**	**5.66**	5.98	9.63	4.95	7.63
GSA	17.00	15.95	12.68	14.37	2.63	15.89	9.89	15.26	8.41	6.58	12.76	7.29	11.56
CNMF	13.46	8.21	10.88	8.26	1.89	7.89	**6.25**	13.69	6.50	5.52	9.77	6.28	8.22
M-FUSE	12.33	9.12	13.82	9.83	2.05	10.95	7.51	10.86	7.29	6.07	11.44	**4.59**	8.82
SASFM	14.62	11.23	14.28	19.10	3.19	14.72	10.83	11.02	10.60	6.44	10.39	8.58	11.25
PNN	19.28	16.74	13.31	14.22	5.52	16.14	15.16	22.18	15.94	11.96	14.83	11.48	14.73
3D-CNN	12.33	10.24	10.13	11.17	2.65	9.15	7.99	14.21	9.13	5.52	9.04	5.96	8.96
ResNet	18.10	15.21	14.86	19.97	5.78	27.09	18.16	23.19	16.40	11.36	11.77	11.80	16.14
MHF-net	9.78	**7.44**	**7.49**	8.86	2.29	**7.20**	7.49	11.13	8.29	**5.10**	**7.18**	5.33	**7.30**
ERGAS results of the 12 testing data (Ideal value: 0)													
FUSE	99.53	147.01	179.22	363.59	114.32	513.42	255.82	102.99	207.90	72.62	127.69	80.55	188.72
ICCV15	94.76	162.98	159.91	308.03	101.64	133.37	60.60	111.35	164.33	74.04	119.79	92.45	131.94
GLP-HS	106.67	143.66	153.45	301.53	82.38	120.59	78.82	88.78	149.76	69.82	119.96	97.01	126.04
SFIM-HS	128.04	162.63	258.07	291.01	62.05	125.61	103.53	90.49	178.70	103.38	163.98	101.41	147.41
GSA	93.73	136.33	144.38	327.03	58.87	108.60	74.08	99.37	163.76	69.07	108.24	86.48	122.50
CNMF	92.62	116.78	175.99	322.04	75.36	100.61	66.00	86.35	138.60	74.38	112.21	104.51	122.12
M-FUSE	109.47	158.44	257.78	283.87	59.95	111.75	137.32	141.99	239.55	78.74	171.98	72.85	151.97
SASFM	208.74	276.10	316.64	1828.96	119.27	403.74	217.30	262.45	193.76	121.77	177.87	225.77	362.70
PNN	108.56	137.69	143.39	302.33	89.35	112.66	104.17	130.42	164.36	107.98	117.76	95.53	134.52
3D-CNN	87.53	121.67	110.53	303.66	57.23	94.73	51.89	134.94	110.68	61.49	78.99	97.09	109.20
ResNet	117.07	133.83	142.78	352.26	145.27	138.57	81.24	121.45	123.55	134.63	78.54	126.14	141.28
MHF-net	**72.06**	**86.51**	**96.13**	**242.24**	**44.94**	**80.37**	**44.82**	**59.59**	**84.93**	**54.67**	**50.82**	**65.40**	**81.87**
SSIM results of the 12 testing data (Ideal value: 1)													
FUSE	0.86	0.83	0.80	0.87	0.84	0.54	0.82	0.86	0.93	0.90	0.91	0.94	0.84
ICCV15	0.91	0.88	0.89	0.87	0.95	0.88	0.97	0.92	0.97	0.92	0.94	0.94	0.92
GLP-HS	0.88	0.81	0.84	0.85	0.90	0.87	0.94	0.90	0.95	0.91	0.90	0.94	0.89
SFIM-HS	0.92	0.89	0.79	0.90	0.95	0.91	0.95	**0.96**	0.96	0.88	0.92	0.94	0.91
GSA	0.85	0.75	0.81	0.84	0.95	0.86	0.96	0.87	0.96	0.91	0.90	0.94	0.88
CNMF	0.93	0.92	0.88	0.91	0.91	0.93	0.97	0.93	0.98	0.89	0.94	0.94	0.93
M-FUSE	0.90	0.88	0.84	0.89	0.95	0.90	0.96	0.93	0.95	0.91	0.90	0.96	0.91
SASFM	0.86	0.76	0.73	0.48	0.81	0.71	0.89	0.93	0.94	0.74	0.87	0.86	0.80
PNN	0.87	0.81	0.89	0.87	0.94	0.89	0.92	0.80	0.92	0.87	0.91	0.92	0.88
3D-CNN	0.92	0.90	0.92	0.91	0.96	0.95	0.98	0.86	0.97	0.94	0.96	0.96	0.94
ResNet	0.81	0.88	0.88	0.82	0.92	0.73	0.92	0.73	0.91	0.89	0.95	0.93	0.86
MHF-net	**0.96**	**0.96**	**0.96**	**0.92**	**0.96**	**0.97**	**0.98**	0.95	**0.98**	**0.95**	**0.98**	**0.97**	**0.96**
FSIM results of the 12 testing data (Ideal value: 1)													
FUSE	0.96	0.94	0.96	0.88	0.94	0.78	0.92	0.97	0.96	0.96	0.96	0.97	0.93
ICCV15	0.97	0.95	0.95	0.89	0.97	0.95	0.98	0.97	0.98	0.96	0.97	0.98	0.96
GLP-HS	0.96	0.91	0.96	0.85	0.94	0.94	0.97	0.94	0.96	0.96	0.96	0.95	0.94
SFIM-HS	0.94	0.91	0.89	0.88	0.96	0.93	0.95	0.96	0.96	0.92	0.93	0.95	0.93
GSA	0.97	0.94	0.96	0.89	0.98	0.96	0.98	0.96	0.98	0.96	0.96	0.97	0.96
CNMF	0.97	0.96	0.95	0.91	0.97	0.97	0.99	0.97	0.98	0.96	0.97	0.97	0.96
M-FUSE	0.96	0.94	0.92	0.88	0.97	0.95	0.97	0.96	0.96	0.95	0.93	0.97	0.95
SASFM	0.94	0.92	0.90	0.76	0.94	0.86	0.95	0.96	0.96	0.93	0.94	0.94	0.92
PNN	0.96	0.96	0.96	0.90	0.96	0.96	0.96	0.95	0.96	0.95	0.96	0.97	0.96
3D-CNN	0.97	0.96	0.98	0.91	**0.98**	0.98	0.99	0.96	**0.98**	**0.98**	0.98	0.98	0.97
ResNet	0.97	0.97	0.97	0.90	0.96	0.98	0.98	0.97	0.97	0.96	0.98	0.97	0.97
MHF-net	**0.98**	**0.98**	**0.98**	**0.92**	0.97	**0.98**	**0.99**	**0.98**	**0.98**	**0.98**	**0.99**	**0.98**	**0.98**

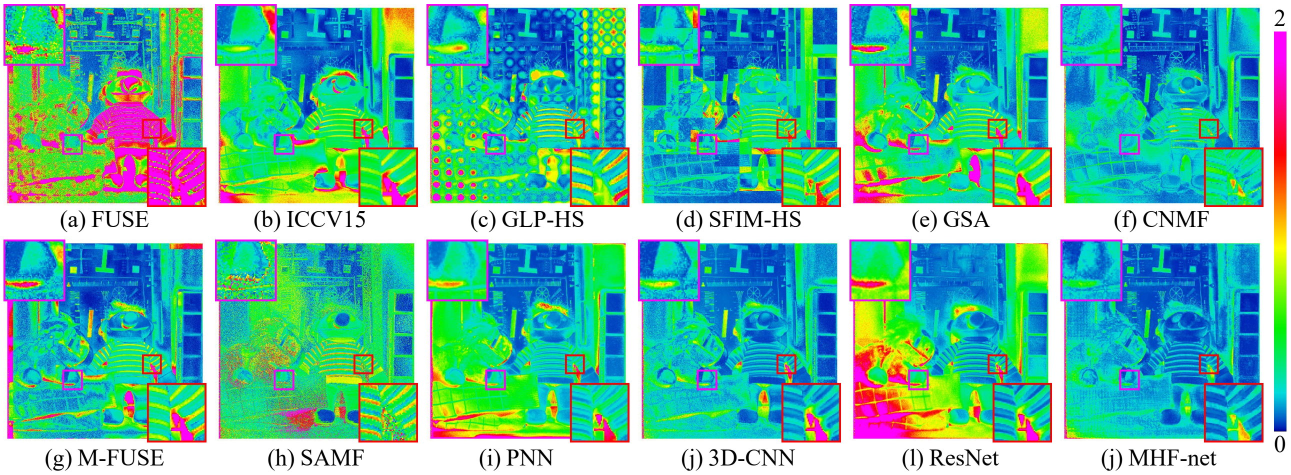


Figure 5. (a)-(h) The error images of the result obtain by the 10 competing method, relative to the reference data, visualized by relative-mean-square error along the spectral mode. Two demarcated areas zoomed in 3 times for easy observation.

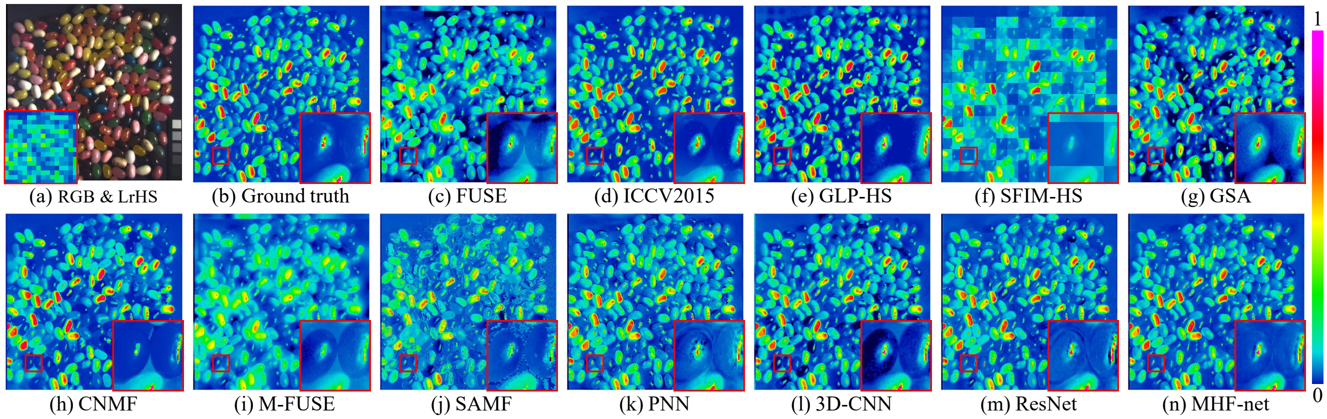


Figure 6. (a) The simulated RGB (HrMS) and LrHS (left bottom) image of *jelly beans*, where we display the 10th (490nm) band of the HS image. (b) The ground-truth HrHS image. (c)-(l) The results obtained by 10 competing methods, with two demarcated areas zoomed in 4 times for easy observation.

same way as the training samples.

Table 1 shows the performance over 12 testing images. It is easy to observe that the proposed method can outperform other methods with respect to all evaluation measures.

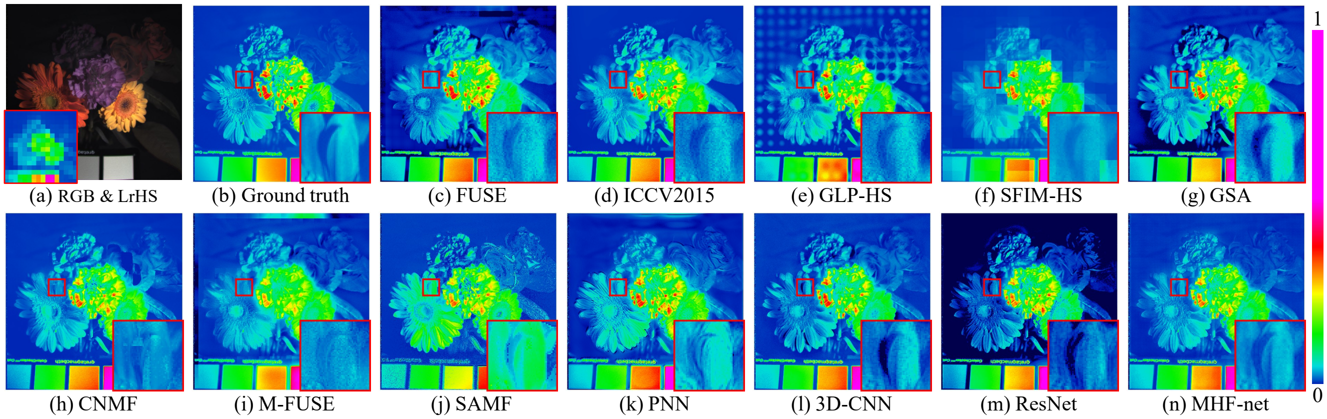
In the main text, we have shown the 10-th band (490nm) of the HS image *chart and staffed toy* obtained by the competing methods visually. Here, we additionally show in Fig. 5 the error images of the result obtain by the 10 competing methods of *chart and staffed toy* relative to the reference data. From the figure, we can easily observe that the error of proposed method is the smallest among all competing methods. To further depict the fusion performance of the proposed method, we show in Fig. 6 - 9 the fusion results of 4 HS images in testing data. From these figures, it is easy to observe that the proposed method performs better than other competing ones, in the better recovery of both the finer-grained textures and the coarser-grained structures.

5.2. More results on Chikusei data

The Chikusei data set [18][9] is an airborne HS imaged taken over Chikusei, Ibaraki, Japan, on 29 July 2014. The data set is of size $2517 \times 2335 \times 128$ with the spectral range from 0.36 to 1.018. We view the original data as the HrHS image and simulate the HrMS (RGB image) and LrMS (with a factor of 32) image in the similar way as the previous section.

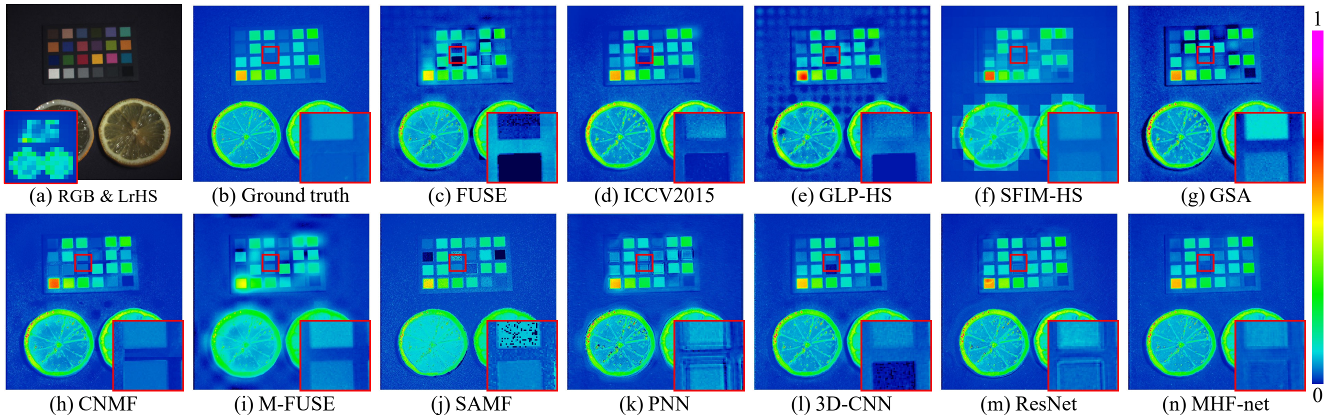
We select a 500×2210-pixel-size image from the top area of the original data to train MHF-net, and extract 96×96 overlapped patches from the training data as reference HrHS images for training. The input HrHS, HrMS and LrHS samples are of size $96 \times 96 \times 128$, $96 \times 96 \times 3$ and $3 \times 3 \times 128$, respectively. Besides, from the remaining part of the original image, we extract 16 non-overlap $448 \times 544 \times 128$ images as testing data. Fig. 10 is an illus-

[9] http://naotoyokoya.com/Download.html


(a) RGB & LrHS (b) Ground truth (c) FUSE (d) ICCV2015 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 7. (a) The simulated RGB (HrMS) and LrHS (left bottom) image of *flowers*, where we display the 10th (490nm) band of the HS image. (b) The ground-truth HrHS image. (c)-(l) The results obtained by 10 competing methods.


(a) RGB & LrHS (b) Ground truth (c) FUSE (d) ICCV2015 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 8. (a) The simulated RGB (HrMS) and LrHS (left bottom) image of *fake and real lemons*, where we display the 10th (490nm) band of the HS image. (b) The ground-truth HrHS image. (c)-(l) The results obtained by 10 competing methods.

tration of the 16 testing images.

Since the large number of spectral bands will highly increase the computational cost, we use the PCA prior in [9] to reduce the computational cost. Specifically, we first compute a $S \times S_r$ matrix $\boldsymbol{V}$ by performing SVD on the HrHS images of the training data:

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{21}$$

where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{HW\times}$ denotes the 500×2210-pixel-size image HrHS images selected as training data, and $\boldsymbol{U} \in \mathbb{R}^{HW\times S_r}$ contains the spectral singular vectors, S_r is the reduced band number, which is set as 30 here and $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values. Then, we compute the following $HW \times S_r$ matrix,

$$\tilde{\boldsymbol{X}}_n = \boldsymbol{X}_n\boldsymbol{V}, \tag{22}$$

where $n = 1, 2, \cdots, N$, N is the sample number. after this, we train our MHF-net with $\left\{(\mathcal{Y}_n, \mathcal{Z}_n), \tilde{\mathcal{X}}_n\right\}_{n=1}^{N}$. Since the channel number of $\tilde{\mathcal{X}}_n$ is much smaller than $\mathcal{X}_n$, the computational cost is thus reduced. When performing testing, we reconstruct the output HrHS image by

$$\hat{\mathcal{X}}_{test} = \text{MHFnet}(\mathcal{Y}_{test}, \mathcal{Z}_{test}, \Theta)\boldsymbol{V}^T. \tag{23}$$

Table 2 shows the performance over 16 testing images. From Table 2, it is easy to observe that the proposed method can outperform other methods with respect to all evaluation measures.

Fig. 11 - 13 shows the composite images of 3 test sample obtained by the competing methods, with bands 70-100-36 as R-G-B. It is easy to observe that the composite images obtained by MHF-net is closest to the ground-truth ones, while the results of other methods usually contain obvious incorrect structure or spectral distortions.

5.3. More results on World View-2 data

In this section, sample images of *Roman Colosseum* acquired by World View-2 (WV-2) is used in our experi-

Table 2. Performance comparison of the competing methods on 16 testing samples in Chikusei data set with respect to 5 PQIs.

Data #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Average
PSNR results of the 16 testing data (Ideal value: $+\infty$)																	
FUSE	20.34	27.39	26.23	24.47	24.78	28.41	23.91	26.57	29.65	29.61	29.87	19.44	26.23	32.38	28.71	27.46	26.59
ICCV15	23.80	27.89	28.13	26.04	26.48	29.06	25.14	28.32	28.72	29.17	29.67	25.09	28.05	31.71	28.77	28.24	27.77
GLP-HS	25.16	28.96	28.94	26.19	27.81	30.74	26.39	28.75	30.85	30.47	30.83	26.07	28.06	32.66	30.02	29.63	28.85
SFIM-HS	24.84	28.64	28.57	26.00	27.54	30.44	26.26	27.85	30.44	29.97	30.32	25.88	28.04	32.25	29.72	29.25	28.50
GSA	22.80	26.53	27.63	25.49	24.73	27.81	24.80	28.37	28.44	28.87	28.85	25.49	26.51	31.12	27.85	27.98	27.08
CNMF	24.86	28.49	28.66	26.52	26.62	31.27	25.31	28.48	30.53	30.65	31.02	26.34	28.51	33.88	29.04	30.30	28.78
M-FUSE	22.63	25.33	25.25	22.91	24.83	27.24	23.07	22.62	26.71	24.71	24.84	22.23	25.25	29.13	25.18	25.60	24.85
SASFM	21.61	26.07	25.92	20.89	24.91	28.00	22.18	26.11	27.35	27.63	25.87	21.72	22.88	28.75	24.82	24.22	24.93
PNN	25.17	27.89	27.34	25.19	27.78	28.09	26.39	27.83	28.96	28.34	26.53	25.61	27.88	28.93	27.89	27.63	27.34
3D-CNN	27.00	30.71	30.89	27.97	29.32	32.20	28.23	31.54	32.70	32.36	31.29	27.46	30.50	34.27	31.41	30.29	30.51
ResNet	27.03	29.79	29.20	27.02	29.17	29.98	28.33	30.09	30.46	30.66	28.93	27.01	30.07	31.41	30.62	29.76	29.35
MHF-net	**28.70**	**33.58**	**32.91**	**29.58**	**31.92**	**34.47**	**30.88**	**32.90**	**34.33**	**32.92**	**31.83**	**29.18**	**32.28**	**34.92**	**33.37**	**32.31**	**32.26**
SAM results of the 16 testing data (Ideal value: 0)																	
FUSE	26.68	6.23	7.73	6.04	12.38	6.28	8.27	6.02	5.26	4.08	4.00	14.47	5.62	2.97	4.37	6.29	7.92
ICCV15	8.92	4.15	3.96	3.14	5.51	3.69	4.48	3.42	3.71	3.20	3.26	3.62	3.29	2.55	3.36	3.51	3.98
GLP-HS	7.67	4.06	3.84	4.07	5.10	3.63	4.95	4.48	3.87	3.43	3.34	3.82	3.58	2.97	3.81	4.05	4.17
SFIM-HS	7.96	4.16	3.85	4.08	5.26	3.64	4.99	4.68	3.96	3.41	3.38	3.79	3.59	2.94	3.87	4.03	4.22
GSA	11.93	5.67	5.21	3.96	8.65	4.88	6.04	4.62	4.79	4.23	4.08	4.32	4.57	3.50	4.96	4.78	5.39
CNMF	5.89	4.15	3.88	3.57	5.15	3.20	4.57	4.07	3.20	3.60	3.44	3.71	2.92	2.61	3.85	3.58	3.84
M-FUSE	10.58	6.61	6.43	6.65	8.85	5.02	7.64	6.91	5.83	5.37	5.44	6.87	5.71	4.07	6.86	7.05	6.62
SASFM	16.47	7.19	7.66	7.55	10.60	6.22	10.86	7.44	7.48	6.75	6.03	7.26	5.76	5.04	6.76	8.16	7.95
PNN	9.15	4.72	4.88	3.99	5.41	4.28	5.22	4.07	4.70	4.15	5.39	4.12	4.36	3.84	4.31	4.17	4.80
3D-CNN	**5.06**	**3.02**	3.16	2.65	3.65	2.90	3.38	2.76	2.70	2.68	**2.90**	2.72	**2.53**	**2.36**	2.87	**2.84**	**3.02**
ResNet	6.53	3.85	3.91	3.07	4.14	3.30	3.93	3.15	3.90	3.18	3.80	2.99	3.54	2.93	3.32	3.45	3.69
MHF-net	6.29	3.13	**2.83**	**2.63**	**3.44**	**2.63**	**3.33**	**2.65**	**2.68**	**2.45**	3.14	**2.55**	2.55	2.44	**2.64**	2.88	**3.02**
ERGAS results of the 16 testing data (Ideal value: 0)																	
FUSE	574.5	207.5	287.0	293.0	310.6	227.7	311.7	257.9	183.7	188.2	198.9	506.6	232.3	140.6	194.6	244.2	272.4
ICCV15	275.3	169.9	170.7	201.0	203.0	160.4	209.6	160.6	163.1	155.7	163.6	216.2	161.8	118.7	154.0	166.6	178.1
GLP-HS	222.1	156.5	154.5	208.5	167.5	141.1	191.7	165.9	136.9	143.6	148.5	192.6	163.3	123.4	148.1	153.5	163.6
SFIM-HS	231.5	164.1	157.0	212.7	173.0	143.5	193.1	176.2	142.1	144.8	150.5	196.7	167.7	123.3	151.0	158.4	167.9
GSA	366.3	247.9	218.2	235.7	309.1	245.5	262.2	210.8	222.3	206.4	230.5	221.6	218.3	178.0	219.7	225.7	238.6
CNMF	232.5	173.9	176.2	199.7	203.3	131.8	230.9	183.8	144.6	158.6	164.0	189.6	153.9	106.0	176.8	148.9	173.4
M-FUSE	308.9	251.1	285.4	337.4	260.3	229.7	309.0	351.5	231.2	297.8	280.4	345.8	245.1	201.5	297.5	279.6	282.0
SASFM	445.9	266.2	335.5	589.4	310.3	253.4	445.5	354.0	284.0	287.8	400.8	439.9	334.1	277.3	422.1	463.5	369.3
PNN	213.6	163.2	175.6	226.0	159.0	168.5	185.4	166.0	148.6	163.2	207.3	200.6	159.4	157.0	168.8	172.9	177.2
3D-CNN	166.3	118.9	122.7	163.7	131.8	112.9	149.2	119.0	104.2	112.3	134.2	161.9	119.0	96.6	120.2	132.9	129.1
ResNet	172.8	135.6	141.9	184.0	134.5	135.6	152.1	134.7	129.5	129.4	165.6	170.6	131.0	117.8	128.9	141.9	144.1
MHF-net	**146.1**	**91.5**	**105.0**	**139.8**	**102.3**	**91.1**	**121.0**	**105.0**	**87.9**	**103.6**	**124.4**	**133.6**	**99.0**	**88.6**	**99.9**	**114.1**	**109.6**
SSIM results of the 16 testing data (Ideal value: 1)																	
FUSE	0.64	0.74	0.72	0.71	0.70	0.71	0.74	0.69	0.75	0.75	0.80	0.55	0.72	0.78	0.77	0.74	0.72
ICCV15	0.66	0.79	0.76	0.75	0.78	0.81	0.78	0.78	0.82	0.81	0.79	0.70	0.84	0.83	0.79	0.77	0.78
GLP-HS	0.79	0.80	0.80	0.74	0.81	0.81	0.79	0.78	0.82	0.81	0.81	0.76	0.81	0.83	0.80	0.79	0.80
SFIM-HS	0.78	0.79	0.80	0.74	0.80	0.81	0.79	0.77	0.81	0.81	0.81	0.75	0.81	0.83	0.80	0.79	0.79
GSA	0.49	0.70	0.66	0.66	0.63	0.72	0.60	0.69	0.74	0.72	0.63	0.67	0.75	0.78	0.67	0.65	0.67
CNMF	0.81	0.77	0.77	0.74	0.77	0.79	0.76	0.72	0.79	0.78	0.79	0.77	0.83	0.83	0.75	0.81	0.78
M-FUSE	0.69	0.66	0.67	0.60	0.69	0.68	0.66	0.49	0.67	0.58	0.66	0.55	0.69	0.69	0.63	0.65	0.64
SASFM	0.69	0.67	0.68	0.57	0.68	0.67	0.67	0.60	0.66	0.65	0.63	0.58	0.62	0.65	0.58	0.57	0.64
PNN	0.82	0.82	0.80	0.77	0.83	0.80	0.81	0.81	0.83	0.81	0.75	0.80	0.83	0.81	0.81	0.81	0.81
3D-CNN	0.88	0.88	0.87	0.82	0.89	0.87	0.86	0.86	0.89	0.87	0.85	0.85	0.89	0.88	0.87	0.85	0.87
ResNet	0.87	0.87	0.87	0.83	0.89	0.85	0.87	0.87	0.87	0.87	0.84	0.86	0.88	0.87	0.88	0.87	0.87
MHF-net	**0.89**	**0.90**	**0.89**	**0.85**	**0.91**	**0.90**	**0.88**	**0.89**	**0.91**	**0.89**	**0.87**	**0.88**	**0.91**	**0.89**	**0.90**	**0.88**	**0.89**
FSIM results of the 16 testing data (Ideal value: 1)																	
FUSE	0.79	0.88	0.85	0.86	0.85	0.86	0.85	0.86	0.89	0.89	0.89	0.75	0.87	0.90	0.89	0.87	0.86
ICCV15	0.79	0.89	0.84	0.83	0.87	0.89	0.88	0.87	0.89	0.88	0.88	0.85	0.89	0.92	0.89	0.88	0.87
GLP-HS	0.90	0.91	0.91	0.87	0.92	0.91	0.90	0.90	0.92	0.90	0.90	0.89	0.91	0.91	0.91	0.90	0.90
SFIM-HS	0.89	0.91	0.90	0.86	0.91	0.91	0.89	0.89	0.92	0.90	0.90	0.88	0.91	0.91	0.90	0.90	0.90
GSA	0.82	0.83	0.83	0.83	0.85	0.85	0.82	0.84	0.86	0.83	0.79	0.82	0.86	0.85	0.84	0.84	0.83
CNMF	0.90	0.90	0.90	0.87	0.90	0.92	0.88	0.88	0.92	0.89	0.89	0.89	0.92	0.93	0.87	0.91	0.90
M-FUSE	0.85	0.86	0.85	0.84	0.86	0.86	0.85	0.83	0.86	0.83	0.85	0.83	0.86	0.86	0.84	0.85	0.85
SASFM	0.84	0.88	0.85	0.81	0.86	0.87	0.83	0.84	0.87	0.85	0.84	0.82	0.84	0.86	0.83	0.83	0.84
PNN	0.91	0.91	0.91	0.89	0.93	0.90	0.91	0.91	0.92	0.91	0.90	0.90	0.91	0.89	0.91	0.91	0.91
3D-CNN	0.93	0.94	0.93	0.91	0.95	0.93	0.93	0.93	0.95	0.93	0.93	0.92	0.94	0.94	0.93	0.92	0.93
ResNet	0.93	0.93	0.93	0.91	0.94	0.92	0.93	0.93	0.93	0.93	0.92	0.92	0.94	0.93	0.94	0.93	0.93
MHF-net	**0.94**	**0.96**	**0.95**	**0.93**	**0.96**	**0.95**	**0.94**	**0.95**	**0.95**	**0.94**	**0.93**	**0.94**	**0.96**	**0.94**	**0.95**	**0.94**	**0.95**

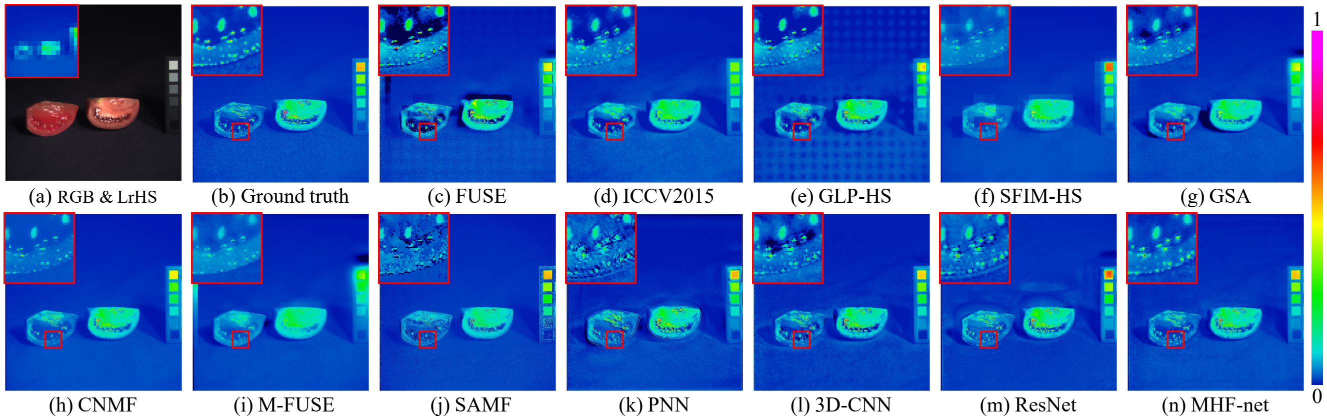


Figure 9. (a) The simulated RGB (HrMS) and LrHS (left top) image of *fake and real tomatoes*, where we display the 10th (490nm) band of the HS image. (b) The ground-truth HrHS image. (c)-(l) The results obtained by 10 competing methods.

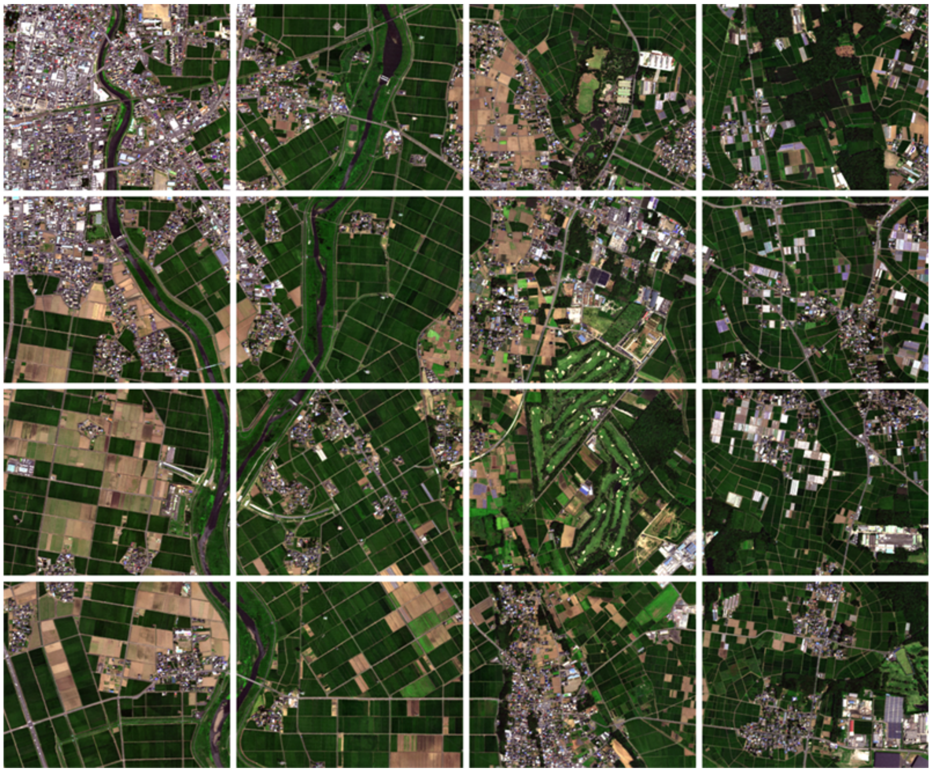
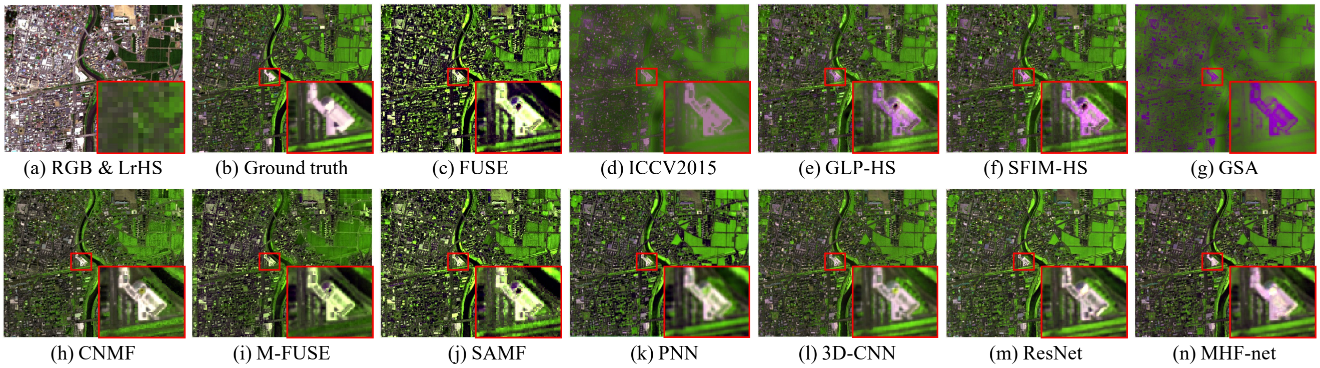


Figure 10. An illustration of the simulated RGB images of the 16 testing samples in Chikusei data.

ments[10] . This data set contains an HrMS image (RGB color image) of size 1676 × 2632 × 3 and an LrHS image of size 419 × 658 × 8, while the HrHS image is unavailable. As shown in Fig. 14, We select the top half part of the HrMS (836 × 2632 × 3) and LrHS (209 × 658 × 8) image to train the MHF-net, and exploit the remaining parts of the data set as testing data. We first extract the training data into 144 × 144 × 3 overlapped HrMS and 36 × 36 × 3 overlapped LrHS patches and then generate the training samples by the method shown in Fig. 3. The input HrHS, HrMS and LrHS samples are of size 36 × 36 × 8, 36 × 36 × 3 and 9 × 9 × 8, respectively.

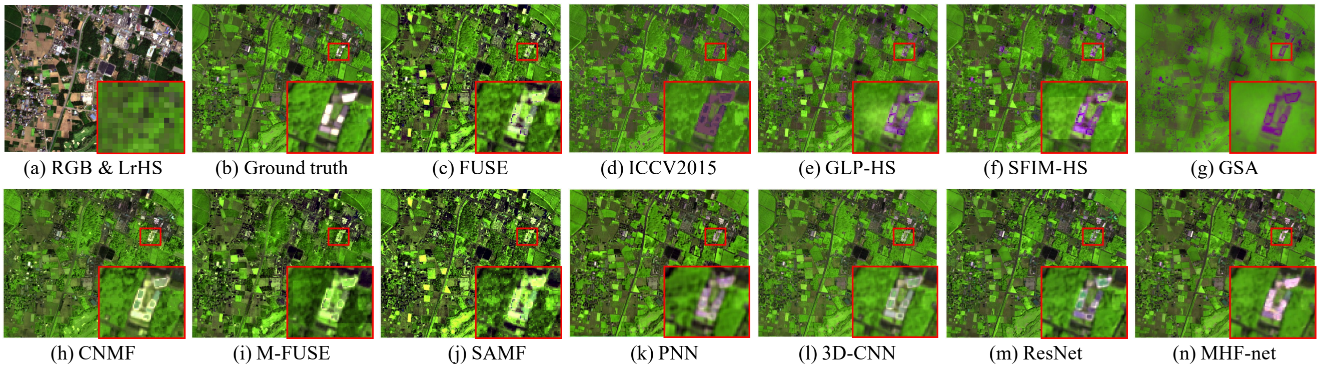
We show in Fig. 15-17 the fusion results of the 3 demarcated area in Fig. 14. Visual inspection evidently shows that the proposed method gives the best result. By comparing the result of ResNet and the proposed method, we can find that the results of these two deep-learning-based methods are both clear, while the color and brightness of result of the proposed method are evidently closer to the LrHS image.

[10] https://www.harrisgeospatial.com/DataImagery/SatelliteImagery/HighResolution/WorldView-2.aspx


(a) RGB & LrHS (b) Ground truth (c) FUSE (d) ICCV2015 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 11. (a) The simulated RGB (HrMS) and LrHS (left bottom) images of a test sample in Chikusei data set, where we show the composite image of the HS image with bands 70-100-36 as R-G-B. (b) The ground-truth HrHS image. (c)-(l) the results obtained by 10 competing methods, with a demarcated area zoomed in 4 times for easy observation.


(a) RGB & LrHS (b) Ground truth (c) FUSE (d) ICCV2015 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 12. (a) The simulated RGB (HrMS) and LrHS (left bottom) images of a test sample in Chikusei data set, where we show the composite image of the HS image with bands 70-100-36 as R-G-B. (b) The ground-truth HrHS image. (c)-(l) the results obtained by 10 competing methods, with a demarcated area zoomed in 4 times for easy observation.

References

[1] B. Aiazzi, S. Baronti, and M. Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007.

[2] N. Akhtar, F. Shafait, and A. Mian. Sparse spatio-spectral representation for hyperspectral image super-resolution. In *European Conference on Computer Vision*, pages 63–78. Springer, 2014.

[3] Boyd, Vandenberghe, and Faybusovich. Convex optimization. *IEEE Transactions on Automatic Control*, 51(11):1859–1859, 2006.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] 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*, 52(3):1693–1704, 2014.

[6] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi. High-resolution hyperspectral imaging via matrix factorization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2329–2336. IEEE, 2011.

[7] C. Lanaras, E. Baltsavias, and K. Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3586–3594, 2015.

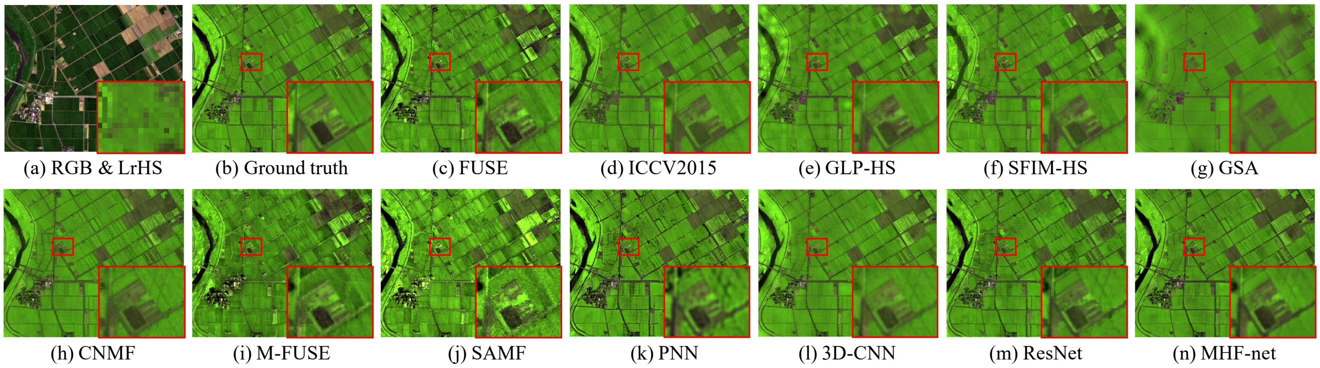
[8] J. Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18):3461–3472, 2000.

[9] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson. Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 14(5):639–643, 2017.

[10] G. Scarpa, S. Vitale, and D. Cozzolino. Target-adaptive cnn-based pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, (99):1–15, 2018.

[11] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti. Hyper-sharpening: A first approach on sim-ga data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):3008–3024, 2015.

[12] L. Wald. *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Presses des lEcole MINES, 2002.


(a) RGB & LrHS (b) Ground truth (c) FUSE (d) ICCV2015 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 13. (a) The simulated RGB (HrMS) and LrHS (left bottom) images of a test sample in Chikusei data set, where we show the composite image of the HS image with bands 70-100-36 as R-G-B. (b) The ground-truth HrHS image. (c)-(l) the results obtained by 10 competing methods, with a demarcated area zoomed in 4 times for easy observation.

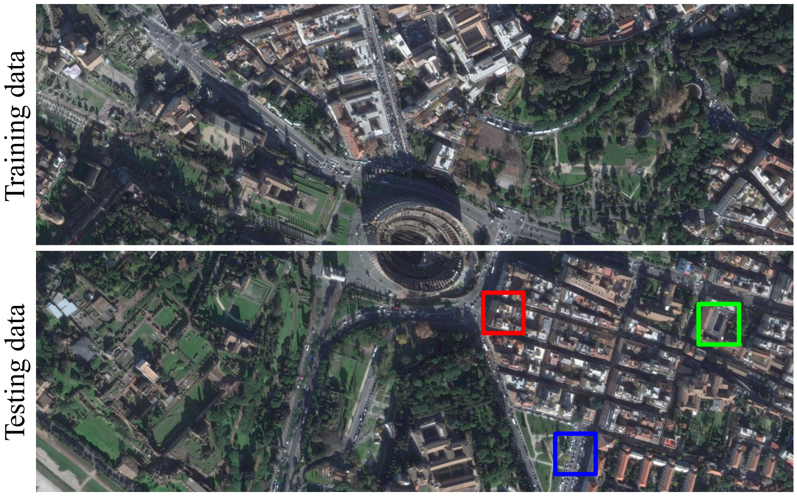


Figure 14. An illustration of RGB image of the World View-2 data. Upper: the training data. Lower: the testing data, where the results of 3 demarcated area will be shown in the later figures.

[13] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

[15] Q. Wei, J. Bioucas-Dias, N. Dobigeon, J.-Y. Tourneret, and S. Godsill. Blind model-based fusion of multi-band and panchromatic images. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2016 IEEE International Conference on*, pages 21–25. IEEE, 2016.

[16] Q. Wei, N. Dobigeon, and J.-Y. Tourneret. Fast fusion of multi-band images based on solving a sylvester equation. *IEEE Transactions on Image Processing*, 24(11):4109–4121, 2015.

[17] 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*, 19(9):2241–2253, 2010.

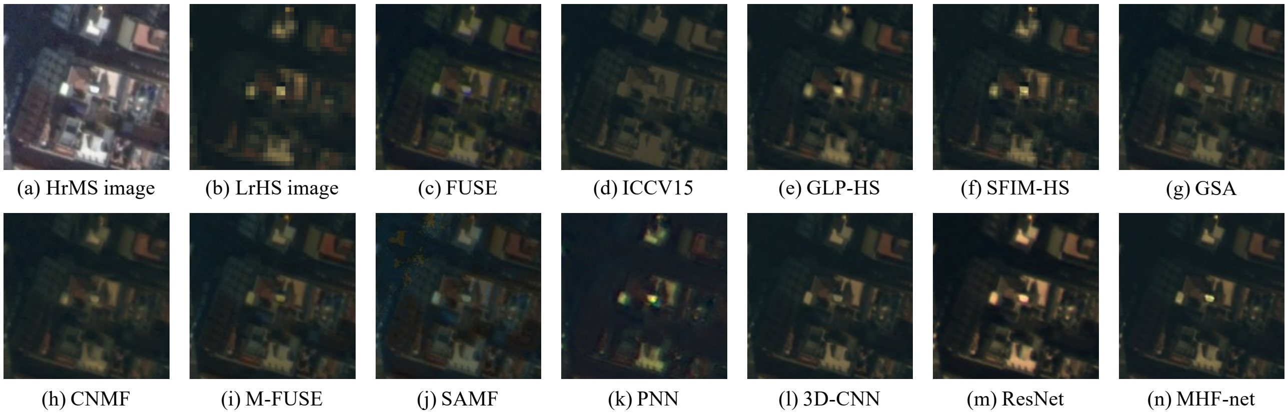
[18] N. Yokoya, C. Grohnfeldt, and J. Chanussot. Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56, 2017.

[19] N. Yokoya, T. Yairi, and A. Iwasaki. Coupled non-negative matrix factorization (CNMF) for hyperspectral and multispectral data fusion: Application to pasture classification. In *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, pages 1779–1782. IEEE, 2011.

[20] R. H. Yuhas, J. W. Boardman, and A. F. Goetz. Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques. 1993.

[21] Y. Zeng, W. Huang, M. Liu, H. Zhang, and B. Zou. Fusion of satellite images in urban area: Assessing the quality of resulting images. In *Geoinformatics, 2010 18th International Conference on*, pages 1–4. IEEE, 2010.

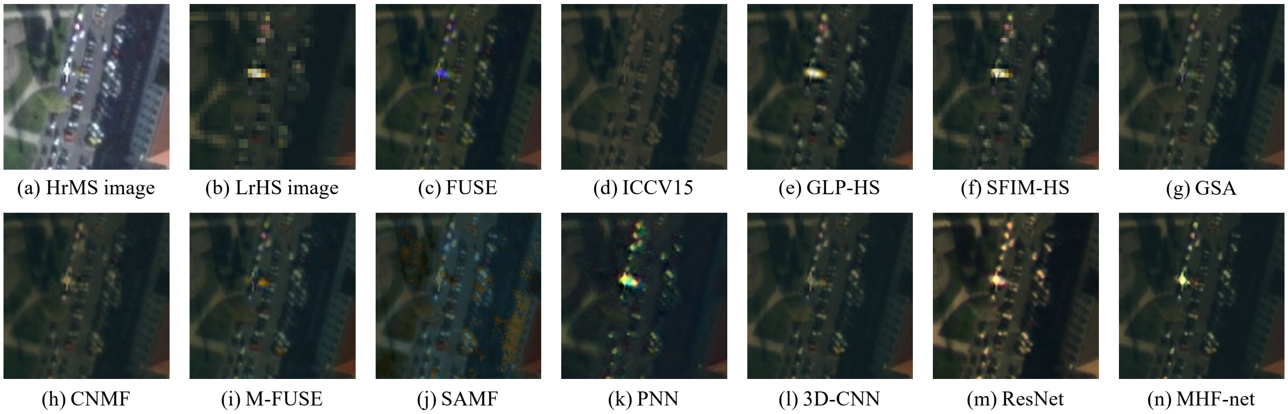
[22] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: a feature similarity index for image quality assessment. *IEEE Trans. Image Processing*, 20(8):2378–2386, 2011.



(a) HrMS image (b) LrHS image (c) FUSE (d) ICCV15 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

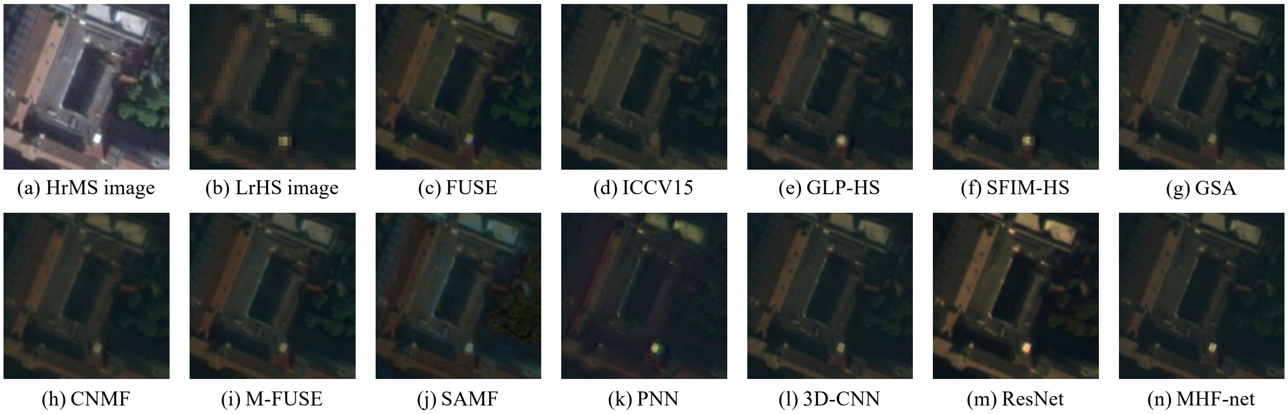
Figure 15. (a) and (b) the HrMS (RGB) and LrHS images of the red demarcated area in Fig. 14, where we show the composite image of the HS image with bands 5-3-2 as R-G-B. (c)-(l) The results obtained by 10 comparison methods.



(a) HrMS image (b) LrHS image (c) FUSE (d) ICCV15 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 16. (a) and (b) the HrMS (RGB) and LrHS images of the blue demarcated area in Fig. 14, where we show the composite image of the HS image with bands 5-3-2 as R-G-B. (c)-(l) The results obtained by 10 comparison methods.



(a) HrMS image (b) LrHS image (c) FUSE (d) ICCV15 (e) GLP-HS (f) SFIM-HS (g) GSA

(h) CNMF (i) M-FUSE (j) SAMF (k) PNN (l) 3D-CNN (m) ResNet (n) MHF-net

Figure 17. (a) and (b) the HrMS (RGB) and LrHS images of the green demarcated area in Fig. 14, where we show the composite image of the HS image with bands 5-3-2 as R-G-B. (c)-(l) The results obtained by 10 comparison methods.