International Journal of Computer Vision (2023) 131:215–242
https://doi.org/10.1007/s11263-022-01699-1

# Memory-Augmented Deep Unfolding Network for Guided Image Super-resolution

Man Zhou[1,2] · Keyu Yan[1,2] · Jinshan Pan[3] · Wenqi Ren[4] · Qi Xie[5] · Xiangyong Cao[6]

Received: 20 January 2022 / Accepted: 22 September 2022 / Published online: 15 October 2022


## Abstract

Guided image super-resolution (GISR) aims to obtain a high-resolution (HR) target image by enhancing the spatial resolution of a low-resolution (LR) target image under the guidance of a HR image. However, previous model-based methods mainly take the entire image as a whole, and assume the prior distribution between the HR target image and the HR guidance image, simply ignoring many non-local common characteristics between them. To alleviate this issue, we firstly propose a maximum a posteriori (MAP) estimation model for GISR with two types of priors on the HR target image, i.e., local implicit prior and global implicit prior. The local implicit prior aims to model the complex relationship between the HR target image and the HR guidance image from a local perspective, and the global implicit prior considers the non-local auto-regression property between the two images from a global perspective. Secondly, we design a novel alternating optimization algorithm to solve this model for GISR. The algorithm is in a concise framework that facilitates to be replicated into commonly used deep network structures. Thirdly, to reduce the information loss across iterative stages, the persistent memory mechanism is introduced to augment the information representation by exploiting the Long short-term memory unit (LSTM) in the image and feature spaces. In this way, a deep network with certain interpretation and high representation ability is built. Extensive experimental results validate the superiority of our method on a variety of GISR tasks, including Pan-sharpening, depth image super-resolution, and MR image super-resolution. Code will be released at https://github.com/manman1995/pansharpening.

**Keywords** Guided image super-resolution · Deep unfolding network · Persistent memory mechanism · Pan-sharpening · Depth image super-resolution · MR image super-resolution

## 1 Introduction

Guided image super-resolution (GISR) has attracted substantial attention and achieved remarkable progress in recent years. Different from the single image super-resolution (SISR) task that only receives input with a single low-resolution (LR) target image, GISR aims to super-resolve the LR target image under the guidance of an additional high-resolution (HR) image whose structural details can help enhance the spatial resolution of the LR target image. Additionally, the guided image in GISR is capable of regularizing the super-resolution process and alleviating the ill-posed problem in SISR, thus leading to better performance.

In GISR, the degradation model is assumed to be the same with SISR, and can be mathematically formulated as

$$\mathbf{L} = (\mathbf{H} * \mathbf{k}) \downarrow_{\mathbf{s}} + \mathbf{n}_{\mathbf{s}} \tag{1}$$

Communicated by Yu Li.

Man Zhou and Keyu Yan have contributed equally to this work.

✉ Xiangyong Cao
caoxiangyong@mail.xjtu.edu.cn

1 Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei, China

2 University of Science and Technology of China, Hefei, China

3 Nanjing University of Science and Technology, Nanjing, China

4 Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

5 School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China

6 School of Computer Science and Technology and Ministry of Education Key Lab For Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, China

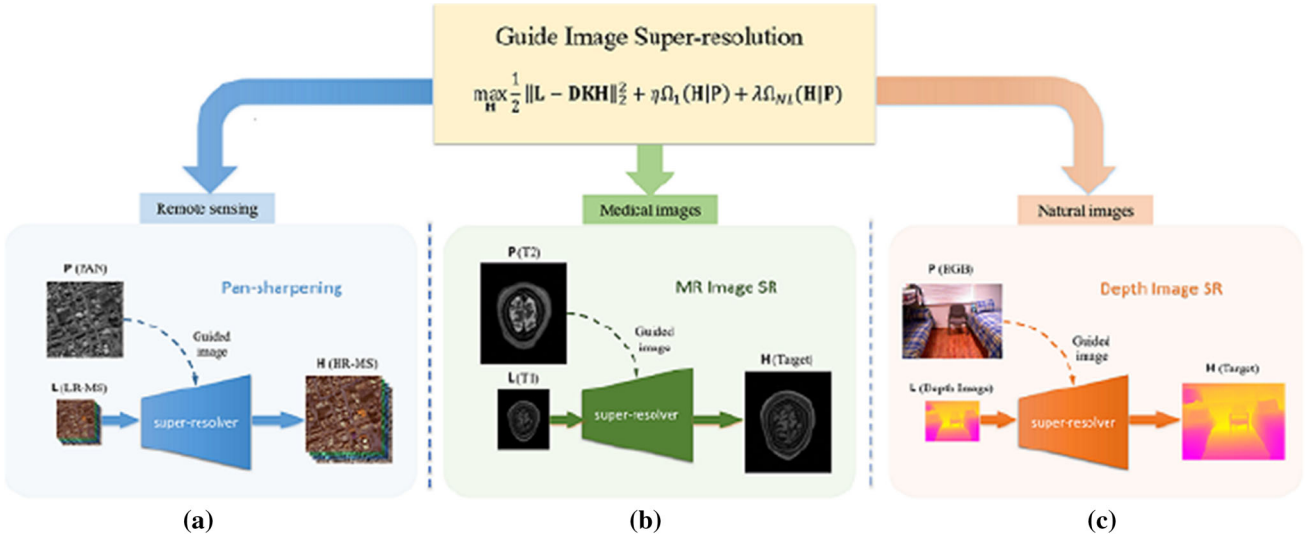



**Fig. 1** Typical Guided Image Super-Resolution (GISR) tasks. **a** Pan-sharpening. **b** MR Image SR. **c** Depth Image SR

where $*$ represents the convolution operation, $\mathbf{L}$ denotes the LR target image, which is obtained by applying the blurring kernel $\mathbf{k}$ and down-sampling operation $\downarrow_{\mathbf{s}}$ on the HR target image $\mathbf{H}$, and $\mathbf{n_s}$ is usually assumed to be additive white Gaussian noise (AWGN) with standard deviation $\sigma$. Different from SISR, GISR uses additional image to guide the super-resolution process. Typical GISR tasks include pan-sharpening (Masi et al., 2016; Xu et al., 2021; Cao et al., 2021), depth image super-resolution (Kim et al., 2021; Su et al., 2019; Sun et al., 2021), and magnetic resonance (MR) image super-resolution (Oktay et al., 2016; Pham et al., 2017; Wang et al., 2020), which are illustrated in Fig. 1. Specifically, pan-sharpening can be considered as a PAN-guided multi-spectral (MS) image SR task, where MS image suffers from the low spatial resolution issue due to the physical hardware limits, and the texture-rich HR PAN image captured from the same scene can act as the guidance image to enhance the spatial resolution of LR MS image. For depth image super-resolution, the HR color image is exploited as a prior for reconstructing regions in depth image, which contains semantically-related and structure-consistent content with the color image. The LR depth image and the HR color image characterize the target and guidance image, respectively. For MR image super-resolution, it aims to super-resolve the MR image of a target contrast under the guidance of the corresponding auxiliary contrast, which provides additional anatomical information. The MR image and the corresponding auxiliary contrast image characterize the target and guidance image, respectively.

Recently, researchers have proposed a large number of guided image super-resolution (GISR) approaches. The main idea of these approaches is transferring structural details of the guidance image to the target image. A large class of methods is filter-based, such as bilateral filtering (Tomasi & Manduchi, 1998; Kopf et al., 2007b), and guided image filtering (He et al., 2012; Shen et al., 2015; Ham et al., 2017), which firstly learns the weights of filter from the guidance image and then applies the learnt weights to filter the target image. These approaches are implemented by hand-crafted objective functions, and may not reflect image priors well. The issue can be alleviated by model-based methods, such as sparsity-based models (Deng & Dragotti, 2019, 2020) and coupled dictionary learning-based models (Wang et al., 2012; Zhuang et al., 2013; Liu et al., 2014; Jing et al., 2015; Bahrampour et al., 2015). These model-based methods aim to capture the correlation among different image modalities by placing explicit hand-crafted priors on different image modalities. Although this type of methods have shown good performance, their priors require careful design and they often require high-computational cost for optimization, limiting their practical applications. Furthermore, the limited representation ability of handcrafted priors leads to unsatisfactory results when processing complex scenes.

Inspired by the success of deep learning in other computer vision tasks, multi-modal deep learning approaches have been developed (Ngiam et al., 2011; Li et al., 2016b; Wu et al., 2018a). A common method is to use a shared layer for fusing different input modalities (Ngiam et al., 2011). Following this idea, a deep neural network (DNN) based joint image filtering method which utilizes the structures of both target and guidance images, was put forward in (Li et al., 2016b). A deep neural network (DNN) reformulation of guided image filtering was designed in (Wu et al., 2018a). However, these DNNs are black-box models in the sense that they neglect the signal structure and properties of the correlation across modalities. To alleviate this black-box issue, some attempts have been

made based on the model-driven deep learning methodology. (Deng & Dragotti, 2019) proposed a deep unfolding network using the deep unfolding design LISTA (Gregor & LeCun, 2010) that solves the sparse coding model. (Marivani et al., 2020; Deng & Dragotti, 2020) proposed multi-modal deep unfolding networks that perform steps similar to an iterative algorithm for convolutional sparse coding model. Nevertheless, current deep unfolding networks (Deng & Dragotti, 2019; Marivani et al., 2020; Deng & Dragotti, 2020) are all built on the (convolutional) spare coding models, where the (convolutional) sparse codes across modalities are assumed to be close. Therefore, these methods only consider some hand-crafted priors and thus do not fully explore the correlation distribution of different modal images. Additionally, the potential of cross-stages for the deep unfolding network has not been fully explored since feature transformation between adjacent stages with reduced channel number leads to information loss. In short, current state-of-the-art deep unfolding methods for GISR suffer from two issues: (1) The correlation distribution of different modal images is not fully exploited, and (2) The unfolding networks suffer from severe information loss in the signal flow. To address the aforementioned issues, this paper first proposes a general GISR model by fully considering the correlation distribution across modalities, and then designs an interpretable information-persistent deep unfolding network by exploring the Long Short-Term Memory mechanisms.

Specifically, we firstly construct a universal variational model for the GISR problem from the maximum a posteriori (MAP) perspective by simultaneously considering two well-established data priors, i.e., local implicit prior and global implicit prior. The local implicit prior models the relationship between the HR target image and the HR guidance image implicitly from a local perspective, and thus can help capture the local related information between the target image and the guidance image. The global implicit prior considers the non-local auto-regression property between the two images from a global perspective, and thus the global correlation between the two images can be well exploited. Since the scene of the target and guidance image in the GISR problem is almost the same, both images thus contain repetitively similar patterns, matching the motivation of the designed non-local auto-regression prior. Secondly, we design a novel alternating optimization algorithm for the proposed variational model, and then unfold the algorithm into a deep network in an effective and transparent manner with cascaded multi-stages. Each stage in the implementation corresponds to three interconnected sub-problems, and each module connects with a specific operator of the iterative algorithm. The detailed flowchart is illustrated in Fig. 2 and the sub-problems are remarked by different colors. In addition, to facilitate the signal flow across iterative stages, the persistent memory mechanism is introduced to augment the information representation by exploiting the Long Short-Term Memory in the image and feature spaces. In the three sub-problems, the output feature maps of each iterative module are selected and integrated for the next iterative stage, thus promoting information fusion across stages and reducing the information loss. In this way, both the interpretation and representation ability of the deep network can be improved. Extensive experimental results validate the superiority of the proposed algorithm against other state-of-the-art methods over the three typical GISR tasks, i.e., pan-sharpening, depth image super-resolution, MR image super-resolution.

In summary, our contributions are four-fold:

- We propose a new GISR model by embedding two image priors, i.e., local implicit prior and global implicit prior, from a maximum a posteriori (MAP) perspective. The two-prior model framework and carefully designed algorithm allow us to separately adopt different types of network modules for the two types of implicit priors when constructing GISR network by adopting deep unrolling technique. Then, we design a novel alternating optimization algorithm for the proposed model.
- We propose an interpretable memory-augmented deep unfolding network (MADUNet) for the GISR problem by unfolding the iterative algorithm into a multistage implementation, which incorporates the advantages of both the model-based prior-equipped methods and data-driven deep-learning methods. With such design, the interpretation of the deep model is improved.
- We propose a new memory mechanism and design a non-local cross-modality module to alleviate the severe information loss issue in the signal flow. The former selectively integrates the features of different-layer and intermediate output of previous stages into the next stage, while the latter explores the information interaction in the target images and across two modalities of the target and guidance images. With such design, the representation ability of the deep model is improved.
- Extensive experiments over three representative GISR tasks, i.e., pan-sharpening, depth image super-resolution and MR image super-resolution demonstrate that our proposed network outperforms other state-of-the-art methods both qualitatively and quantitatively.

## 2 Related Work

### 2.1 Single Image Super-resolution

Single image super-resolution (SISR) methods have been extensively studied in recent decades, and these methods can be roughly divided into three categories, i.e., interpolation-

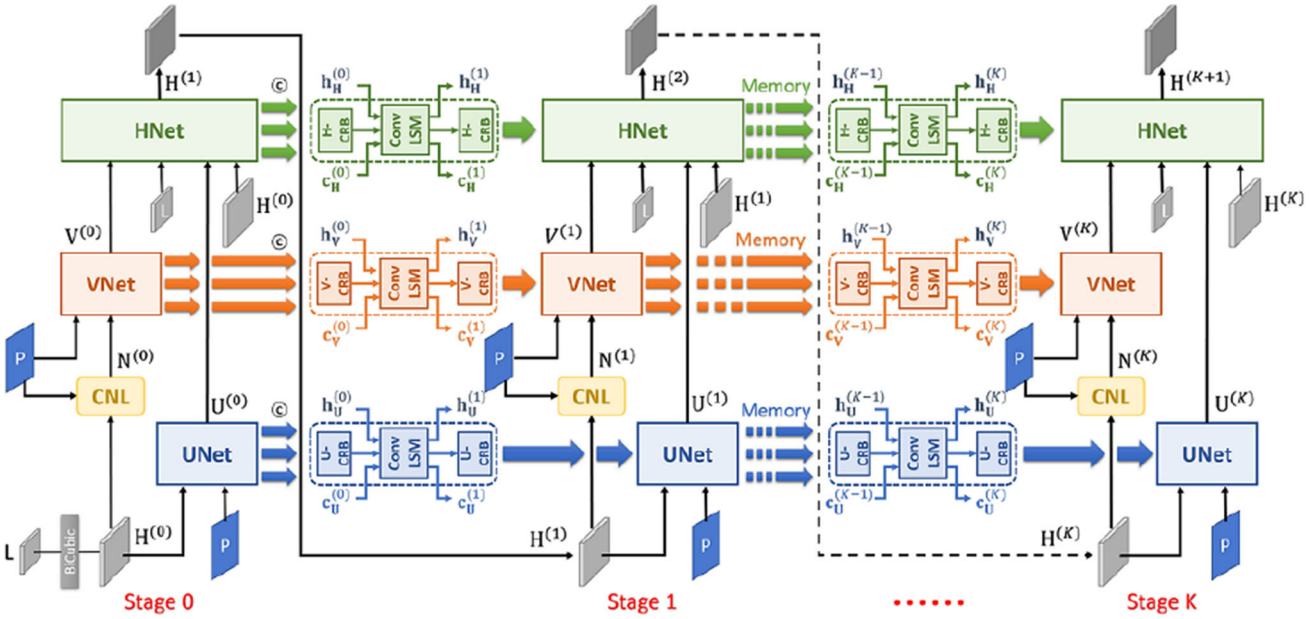



**Fig. 2** The overall architecture of our proposed unfolding network, consisting of information flow and memory flow. For information flow, the LR target image **L** is firstly up-sampled as $\mathbf{H}^{(0)}$ and then performs the stage-wise iteration updating in the overall K stages where UNet and VNet are parallel updated and then transmitted into HNet. To emphasize, CNL module aims to explore the cross-modality information to transfer the detailed structures of guidance image into the target image to generate the immediate output **N**. To facilitate the signal flow across iterative stages, the persistent memory mechanism (remarked as memory flow) is introduced to augment the information representation by exploiting the Long short-term memory unit (LSTM) in the image and feature spaces

based approaches, reconstruction-based approaches, and learning-based approaches.

Interpolation-based methods (Dai et al., 2007; Sun et al., 2008; Sanchez-Beato & Pajares, 2008) are easy to implement, but tend to overly smooth image edges and bring in aliasing and blurring effects. Reconstruction-based methods (Mallat & Yu, 2010; Dong et al., 2012; Yang et al., 2013) utilize various image priors to regularize the ill-posed SISR problem, and always outperform the interpolation-based methods. Learning-based methods (Yang et al., 2008, 2010, 2012; Jia et al., 2012; Timofte et al., 2013, 2017; Dong et al., 2015; Kim et al., 2016a, b; Mao et al., 2016; Bruna et al., 2015; Zhang et al., 2018a; Tai et al., 2017; Zhang et al., 2018b) are implemented by directly learning the mapping function from LR image to HR image using some machine learning techniques. For example, several sparse representation and dictionary learning based methods were proposed in (Yang et al., 2008, 2010, 2012; Jia et al., 2012), which assume that HR/LR image pairs share the same sparse coefficients in terms of a pair of HR and LR dictionaries. Later, several anchored neighbourhood regression based methods were developed in (Timofte et al., 2013, 2017), which combine the neighbor embedding technique and dictionary learning method. Recently, deep learning methods have become the most popular technique among the learning-based methods for SISR. SRCNN (Dong et al., 2015) was the first deep convolutional network for SISR. A faster version FSRCNN was designed in (Dong et al., 2016). Subsequently, a series of SISR deep learning methods focus on improving the network complexity by increasing the depth of the network. For example, a 20-layer residual network for SR was proposed in (Kim et al., 2016a, b), a 30-layer convolutional autoencoder was developed in (Mao et al., 2016), and a 52-layer deep recursive residual network was designed in (Tai et al., 2017). Additionally, residual learning has been used to learn inter-layer dependencies in (Zhang et al., 2018a, b). Deep unfolding technique has also been applied to SISR in (Liu et al., 2016; Zhang et al., 2020), where the iterative algorithm is implemented by a neural network.

## 2.2 Guided Image Super-resolution

Compared with single image SR, guided image SR utilizes an additional HR guidance image to super-resolve the LR target image by transferring structural information of the guidance image to the target image.

The bilateral filter (Tomasi & Manduchi, 1998) is essentially a translation-variant edge-preserving filter, which calculates a pixel value by the weighted average of its neighboring pixels, and the weights are learnt from the given image using filter kernels. The joint bilateral upsampling (Kopf et al., 2007b) is a generalization of the bilateral filter. Instead

of computing the weights from the input image, the joint bilateral upsampling learns the weight from an additional guidance image using a range filter kernel. In this way, the high frequency components of the guidance image can be transferred into the LR target image. However, this method may introduce gradient reversal artifacts. To mitigate this issue, guided image filtering (He et al., 2012) was designed by directly transferring the guidance gradients into the LR target image. But this method still brings in notable appearance change. To alleviate this issue, (Shen et al., 2015) put forward a model to optimize a novel scale map for capturing the nature of structure discrepancy between the target and guidance images. However, since the above methods only consider the static guidance image, these methods may convert inconsistent structure and detail information of guidance image to the target image. (Ham et al., 2017) developed a robust guided image filtering method to iteratively refine the target image by utilizing the static guidance image and the dynamic target image. (Song et al., 2019) proposed a coupled dictionary learning based GISR method, which learns a set of dictionaries that couple different image modalities in the sparse feature domain from a training dataset, and thus alleviate the inconsistency issue between the guidance and target images. These approaches adopt hand-crafted objective functions, and may not reflect natural image priors well.

Recently, deep learning-based methods have been applied for this problem. A deep neural network (DNN) based joint image filtering method which simutaneously utlizes the structures of both target and guidance images, was put forward in (Li et al., 2016b). A deep neural network (DNN) reformulation of guided image filtering was designed in (Wu et al., 2018a). A deep neural network (DNN) by exploiting a customized transformer architecture was devised in (Zhou et al., 2022a). Further, a novel and effective method is proposed by exploiting a customized transformer architecture and information-lossless invertible neural module for long-range dependencies modeling and effective feature fusion(Zhou et al., 2022b). A novel mutual information-driven framework is proposed in (Zhou et al., 2022c) to explicitly enforce the complementary information learning. The four methods are both pure data-driven methods and rely on large amount of image pairs to train the network. Additionally, based on the model-driven deep learning methodology, (Marivani et al., 2020) proposed a multimodal deep unfolding network that performs steps similar to an iterative algorithm for convolutional sparse coding with side information. (Deng & Dragotti, 2019) proposed another deep unfolding network for guided image SR using the deep unfolding design LISTA (Gregor & LeCun, 2010), where the latent representations of the input images are computed by two LISTA branches, and the HR target image is finally generated by the linear combination of these representations. Later, similar method was proposed in (Deng & Dragotti, 2020) by using three LISTA branches to separate the common feature shared by different modalities and unique features for each modality, and the target image is reconstructed by the combination of these common and unique feature. Almost all the deep learning based GISR methods only consider the local features since these methods only employ the convolution neural operators to construct the model architectures, and only this work (Zhou et al., 2022b) proposed by our team considers both features. We proposed a novel pan-sharpening method (Zhou et al., 2022b) by combining the local feature modeling of CNN and the global feature modeling of customized multi-modality transformer. Differently, we propose a new memory-augmented GISR model by embedding two image priors, i.e., local implicit prior and global implicit prior, from a maximum a posterior (MAP) perspective.

The most relevant work to our proposed method is (Song et al., 2021). Compared with (Song et al., 2021), our work mainly differs from the following aspects. Firstly, our work targets at the guided image super-resolution application which is essentially a multi-modal image processing task, and aims to super-resolve the LR target image under the guidance of an additional high-resolution (HR) image, while (Song et al., 2021) only focuses on the compressive sensing of single-modal image. Secondly, although our work and (Song et al., 2021) both consider the memory-augmented mechanism, their characteristics are fundamentally different. The memory flow of (Song et al., 2021) is coupled with the signal flow, which may result in the information aliasing issue. Differently, our devised memory mechanism is multiple independent memory flows that are orthogonal to signal flow, and thus our proposed memory mechanism can alleviate the information aliasing issue in some sense. Additionally, due to the orthogonal property of memory flow and signal flow, our proposed network is still well defined after the memory flow is removed. As for (Song et al., 2021), if the memory flow is deleted, the network structure of (Song et al., 2021) will break down since the memory flow and signal flow are coupled together. In conclusion, the memory-augmented mechanism of our method is more effective and reasonable.

## 3 Proposed Approach

In this section, we provide a detailed introduction to our proposed memory-augmented deep unfolding GISR network, illustrated in Fig. 2. For convenience, we first define some notations used in the GISR model. To be specific, $\mathbf{L} \in \mathbb{R}^{m\times n\times B}$ denotes the low-resolution (LR) target image, $\mathbf{H} \in \mathbb{R}^{M\times N\times B}$ represents the corresponding high-resolution (HR) target image, and $\mathbf{P} \in \mathbb{R}^{M\times N\times b}$ is the guidance image.

### 3.1 MAP Model for Guided Image Super-resolution

In GISR, we assume that the LR target image $\mathbf{L}$ is obtained through performing the blurring kernel $\mathbf{k}$ and down-sampling operator over the HR target image $\mathbf{H}$, and thus the degradation model can be mathematically formulated as

$$\mathbf{L} = (\mathbf{H} * \mathbf{k}) \downarrow_{\mathbf{s}} + \mathbf{n}_{\mathbf{s}}, \tag{2}$$

where $*$ represents the convolution operation, and $\mathbf{n}_{\mathbf{s}}$ is usually assumed to be additive white Gaussian noise (AWGN) with standard deviation $\sigma$. The spatial resolution ratio between $\mathbf{H}$ and $\mathbf{L}$ is $r = M/m = N/n$. The observation model in Eq. (2) can be equivalently reformulated as

$$\mathbf{L} = \mathbf{DKH} + \mathbf{n}_{\mathbf{s}}, \tag{3}$$

where $\mathbf{K}$ is the matrix form of kernel $\mathbf{k}$, and $\mathbf{D}$ is the matrix form of down-sampling operator. Based on the observation model in Eq. (3), the distribution of $\mathbf{L}$ is defined as

$$P(\mathbf{L}|\mathbf{H}) = \mathcal{N}(\mathbf{L}|\mathbf{DKH}, \sigma^2\mathbf{I}), \tag{4}$$

where $\mathcal{N}(\mathbf{L}|\mathbf{DKH}, \sigma^2\mathbf{I})$ denotes the Gaussian distribution with mean $\mathbf{DKH}$ and covariance matrix $\sigma^2\mathbf{I}$.

Since GISR super-resolves the LR target image under the guidance image, which is usually captured in the same scene with the LR target image, the LR target image and the guidance image thus share some global and local relevant features. To capture both features, we design two types of image priors, i.e., local implicit prior and global implicit prior. The local implicit prior models the relationship between the HR target image and the HR guidance image implicitly from a local perspective, and thus can help capture the local related information between the target image and the guidance image. The global implicit prior considers the non-local auto-regression property between the two images from a global perspective, and thus the global correlation between the two images can be well exploited. The non-local auto-regressive property essentially reflects the image self-similarity, and it constrains the image local structure (i.e., the local patch) by using the non-local redundancy (Dong et al., 2013). This idea is first applied in the natural image interpolation (Dong et al., 2013) which assumes that many non-local similar patches to a given patch could provide non-local constraint to the local structure. Since the target and guidance image capture the same scene, consistent patterns thus exist between the two images. The HR target image can be reconstructed by aggregating the long-range correlation information from both target and guidance images as shown in Fig. 3.

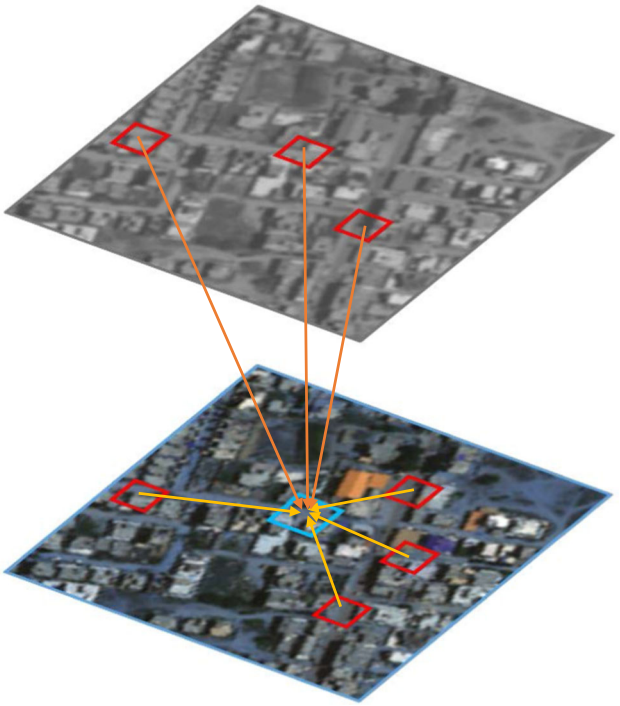


**Fig. 3** Illustration of non-local cross-modality aggregation. Since the target and guidance image capture the same scene, consistent patterns thus exist between the two images. The HR target image can be reconstructed by aggregating the long-range correlation information from both target and guidance images

Specifically, we assume the local implicit prior distribution $P_1(\mathbf{H}|\mathbf{P})$ and global implicit prior distribution $P_2(\mathbf{H}|\mathbf{P})$ separately as follows:

$$P_1(\mathbf{H}|\mathbf{P}) \propto \exp\{-\eta\Omega_1(\mathbf{H}|\mathbf{P})\}, \tag{5}$$
$$P_2(\mathbf{H}|\mathbf{P}) \propto \exp\{-\lambda\Omega_{NL}(\mathbf{H}|\mathbf{P})\}, \tag{6}$$

where $\Omega_1(\mathbf{H}|\mathbf{P})$ and $\Omega_{NL}(\mathbf{H}|\mathbf{P})$ are two energy functions related to $\mathbf{H}$ and $\mathbf{P}$, $\eta$ and $\lambda$ are the weight parameters. For simplicity, we assume the distribution of $\mathbf{H}$ is

$$P(\mathbf{H}|\mathbf{P}) \propto P_1(\mathbf{H}|\mathbf{P})P_2(\mathbf{H}|\mathbf{P}), \tag{7}$$

Therefore, the posterior of $\mathbf{H}$ given $\mathbf{L}$ and $\mathbf{P}$ can be computed by the Bayes formula:

$$P(\mathbf{H}|\mathbf{L},\mathbf{P}) = \frac{P(\mathbf{L}|\mathbf{H})P(\mathbf{H}|\mathbf{P})}{P(\mathbf{L}|\mathbf{P})}, \tag{8}$$

where $P(\mathbf{L}|\mathbf{P})$ is the marginal distribution of $\mathbf{L}$ which is not related with $\mathbf{H}$. By using the maximum a posterior (MAP) principle, $\mathbf{H}$ can be obtained by maximizing the log-posterior $\log P(\mathbf{H}|\mathbf{L},\mathbf{P})$, which is equivalent to the following optimization problem:

$$\max_{\mathbf{H}} \log P(\mathbf{L}|\mathbf{H}) + \log P_1(\mathbf{H}|\mathbf{P}) + \log P_2(\mathbf{H}|\mathbf{P}). \tag{9}$$

Further, Eq. (9) can be reformulated as

$$\max_{\mathbf{H}} \frac{1}{2}||\mathbf{L}-\mathbf{DKH}||_2^2+\eta\Omega_1(\mathbf{H}|\mathbf{P})+\lambda\Omega_{NL}(\mathbf{H}|\mathbf{P}). \quad (10)$$

Eq. (10) is our final proposed model. In the next section, we will develop an optimization algorithm to solve this model.

### 3.2 Model Optimization

We now solve the optimization problem of Eq. (10) using half-quadratic splitting (HQS) algorithm, which has been widely used in solving image inverse problems (Geman & Reynolds, 1992; Geman & Yang, 1995; Krishnan & Fergus, 2009; He et al., 2014). By introducing two auxiliary variables $\mathbf{U}$ and $\mathbf{V}$, Eq. (10) can be reformulated as a non-constrained optimization problem:

$$\min_{\mathbf{H},\mathbf{U},\mathbf{V}} \frac{1}{2}||\mathbf{L}-\mathbf{DKH}||_2^2+\frac{\eta_1}{2}||\mathbf{U}-\mathbf{H}||_2^2+\eta\Omega_1(\mathbf{U}|\mathbf{P}) + \frac{\lambda_1}{2}||\mathbf{V}-\mathbf{H}||_2^2+\lambda\Omega_{NL}(\mathbf{V}|\mathbf{P}) \quad (11)$$

where $\eta_1$ and $\lambda_1$ are penalty parameters. When $\eta_1$ and $\lambda_1$ simultaneously approach to infinity, the solution of minimizing Eq. (11) converges to that of minimizing Eq. (10). Then, minimizing Eq. (11) can be achieved by solving three subproblems for alternately updating $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{H}$.

**Updating U**. Given the estimated HR target image $\mathbf{H}^{(k)}$ at iteration $k$, the auxiliary variable $\mathbf{U}$ can be updated as:

$$\mathbf{U}^{(k)}=\arg\min_{\mathbf{U}} \frac{\eta_1}{2}\left|\left|\mathbf{U}-\mathbf{H}^{(k)}\right|\right|_2^2+\eta_2\Omega_1(\mathbf{U}|\mathbf{P}). \quad (12)$$

By applying the proximal gradient method (Rockafellar, 1976) to Eq. (12), we can derive

$$\mathbf{U}^{(k)}=\text{prox}_{\Omega_1(.)}(\mathbf{U}^{(k-1)}-\delta_1\nabla f_1(\mathbf{U}^{(k-1)})) \quad (13)$$

where $\text{prox}_{\Omega_1}(\cdot)$ is the proximal operator corresponding to the implicit prior $\Omega_1(\cdot)$, $\delta_1$ denotes the updating step size, and the gradient $\nabla f_1(\mathbf{U}^{(k-1)})$ is

$$\nabla f_1(\mathbf{U}^{(k-1)})=\mathbf{U}^{(k-1)}-\mathbf{H}^{(k)}. \quad (14)$$

**Updating V**. Given $\mathbf{H}^{(k)}$, $\mathbf{V}$ can be updated as:

$$\mathbf{V}^{(k)}=\arg\min_{\mathbf{V}} \frac{\lambda_1}{2}\left|\left|\mathbf{V}-\mathbf{H}^{(k)}\right|\right|_2^2+\lambda\Omega_{NL}(\mathbf{V}|\mathbf{P}). \quad (15)$$

Similarly, we can obtain

$$\mathbf{V}^{(k)}=\text{prox}_{\Omega_{NL}(.)}(\mathbf{V}^{(k-1)}-\delta_2\nabla f_2(\mathbf{V}^{(k-1)})) \quad (16)$$

where $\text{prox}_{\Omega_{NL}}(\cdot)$ is the proximal operator corresponding to the non-local prior term $\Omega_{NL}(\cdot)$, $\delta_2$ indicates the updating step size, and the gradient $\nabla f_2(\mathbf{V}^{(k-1)})$ is computed as

$$\nabla f_2(\mathbf{V}^{(k-1)})=\mathbf{V}^{(k-1)}-\mathbf{H}^{(k)}. \quad (17)$$

**Updating H**. Given $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$, $\mathbf{H}$ is updated as:

$$\mathbf{H}^{(k+1)}=\arg\min_{\mathbf{H}} \frac{1}{2}||\mathbf{L}-\mathbf{DKH}||_2^2+\frac{\eta_1}{2}\left|\left|\mathbf{U}^{(k)}-\mathbf{H}\right|\right|_2^2 + \frac{\lambda_1}{2}\left|\left|\mathbf{V}^{(k)}-\mathbf{H}\right|\right|_2^2. \quad (18)$$

We can derive closed form solution of updating $\mathbf{H}$ from Eq. (18)

$$\mathbf{H}^{(k+1)}=\left((\mathbf{DK})^T\mathbf{DK}+\frac{\eta_1}{2}\mathbf{I}+\frac{\lambda_1}{2}\mathbf{I}\right)^{-1}\left((\mathbf{DK})^T\mathbf{L}+\frac{\eta_1}{2}\mathbf{U^k}+\frac{\lambda_1}{2}\mathbf{H}\right) \quad (19)$$

and the above updating equation relies on calculating the inverse of a large matrix, which is computational inefficiency. To alleviate this issue, we still follow the updating rules of $\mathbf{U}$ and $\mathbf{V}$ and adopt the gradient decent method to update $\mathbf{H}$. Therefore, the updating equation for $\mathbf{H}$ is

$$\mathbf{H}^{(k+1)}=\mathbf{H}^{(k)}-\delta_3\nabla f_3(\mathbf{H}^{(k)}), \quad (20)$$

where $\delta_3$ is the step size, and the gradient $\nabla f_3(\mathbf{H}^{(k)})$ is

$$\nabla f_3(\mathbf{H}^{(k)})=(\mathbf{DK})^T(\mathbf{DKH}^{(k)}-\mathbf{L})+\eta_1(\mathbf{H}^{(k)}-\mathbf{U}^{(k)}) + \lambda_1(\mathbf{H}^{(k)}-\mathbf{V}^{(k)}) \quad (21)$$

where $T$ is the matrix transpose operation.

### 3.3 Deep unfolding network

Based on the iterative algorithm, we build a deep neural network for GISR as illustrated in Fig. 2. This network is an implementation of the algorithm for solving Eq. (10). Each stage receives the inputs $\mathbf{U}^{(k-1)}$, $\mathbf{V}^{(k-1)}$, and $\mathbf{H}^{(k)}$, and generates the outputs $\mathbf{U}^{(k)}$, $\mathbf{V}^{(k)}$, and $\mathbf{H}^{(k+1)}$.

In the proposed algorithm, the two proximal operators $\text{prox}_{\Omega_1}(\cdot)$ and $\text{prox}_{\Omega_{NL}}(\cdot)$ can not be explicitly deduced since the regularization terms $\Omega_1(\cdot)$ and $\Omega_{NL}(\cdot)$ are not explicitly defined. We thus use deep CNNs to learn the two proximal operators for updating $\mathbf{U}^{(k-1)}$ and $\mathbf{V}^{(k-1)}$.

Specifically, we can easily learn $\text{prox}_{\Omega_1}(\cdot)$ in Eq. (13) using the CNN, dubbed as UNet as follows:

$$\mathbf{U}^{(k)}=\text{prox}_{\Omega_1}(\mathbf{U}^{(k-1)},\mathbf{H}^{(k)}) \approx \text{UNet}(\mathbf{U}^{(k-1)},\mathbf{H}^{(k)},\mathbf{P}), \quad (22)$$

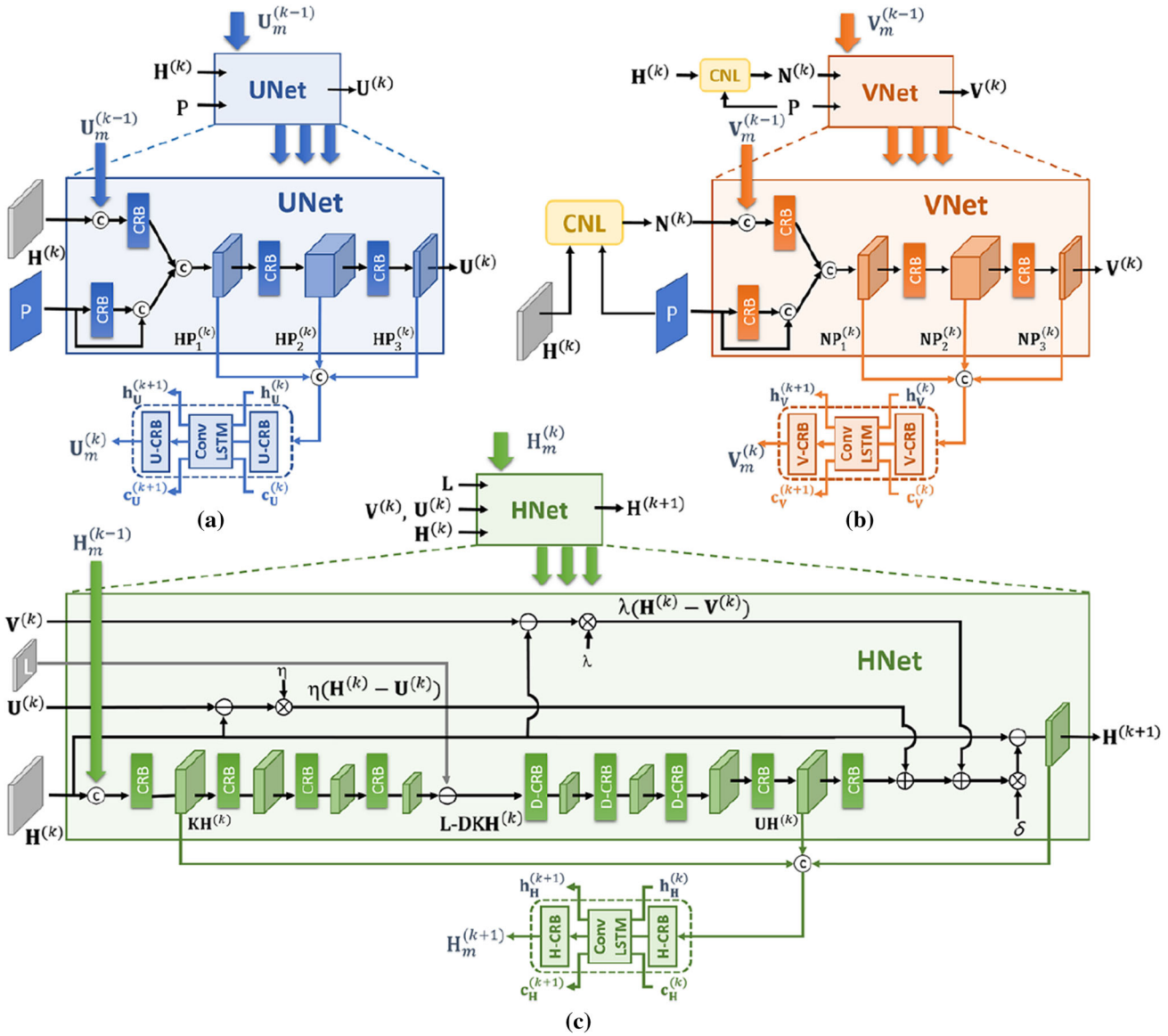



**Fig. 4** The detailed architecture of UNet, VNet, and HNet in stage $k$ of the overall network. Each module corresponds to one operation of the iterative step for updating **U**, **V**, and **H**. The basic CRB unit consists of the pure convolution layer and the effective residual blocks. The structures of CRB, V − CRB, U − CRB and H − CRB are the same but with different channel settings in transformation

where **P** is added as an input of the UNet due to the regularization term $\Omega_1(\mathbf{U}|\mathbf{P})$ in Eq. (12) that implies **U** and **P** have some relations. UNet represents the CNN whose input and output have the same spatial resolution. The detailed implementation of UNet is illustrated in Fig. 4a.

Before learning $\text{prox}_{\Omega_{NL}}(\cdot)$ using the CNN, we firstly introduce an intermediate variable $\mathbf{N}^{(k)}$ to explicitly consider the non-local term $\Omega_{NL}(\mathbf{H}|\mathbf{P})$ by devising a cross-modalities non-local module, denoted as CNL, which is illustrated in Fig. 5 and takes $\mathbf{H}^{(k)}$ and **P** as input and then generates $\mathbf{N}^{(k)}$. The motivations of coming up with the special CNL structure are shown as follows. Firstly, for the guided image super-resolution task (i.e., pan-sharpening), its non-local auto-regressive property mainly lies in two aspects, namely the non-local property between the HR target image H and itself, and the non-local property between the HR target image H and the HR guidance image P. Secondly, inspired by the structure of the non-local neural network (Wang et al., 2018), we thus design the structure of the left branch in Fig. 5 to model the non-local property between H and itself and the structure of the right branch in Fig. 5 to model the non-local property between H and P. Therefore, based on the above two reasons, we come up with this special structure for the guided

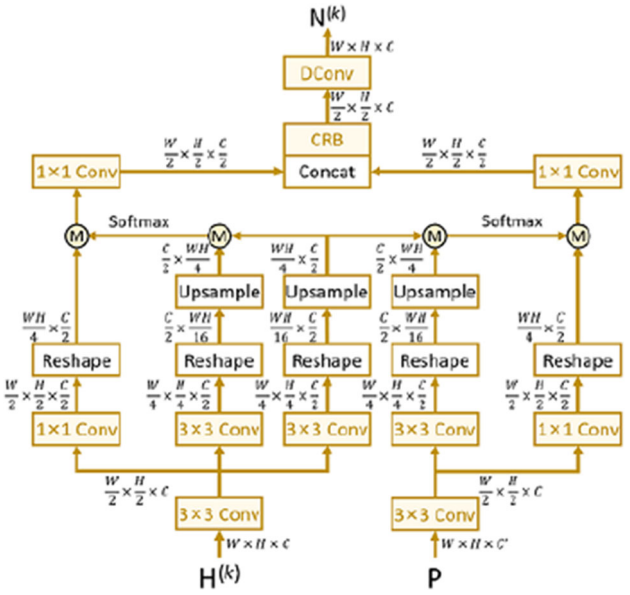



**Fig. 5** The cross-modality non-local operation module. It takes the updated HR target image $\mathbf{H}^{(k)}$ and guidance image $\mathbf{P}$ as input and generates the refined image $\mathbf{N}^{(k)}$, referred to Eq. (23). M represents matrix multiplication

image super-resolution task. The CNL module is defined as

$$\mathbf{N}^{(k)} = \mathrm{CNL}(\mathbf{H}^{(k)}, \mathbf{P}). \tag{23}$$

The output $\mathbf{N}^{(k)}$ of CNL module can be interpreted as a slight fine-tuning on the HR target image $\mathbf{H}^{(k)}$ using the guided image $\mathbf{P}$. Then, $\mathbf{N}^{(k)}$ is plugged into Eq. (17) to replace $\mathbf{H}^{(k)}$ for computing the gradient $\nabla f_2(\mathbf{V}^{(k-1)})$. Finally, we can also learn $\mathrm{prox}_{\Omega_{NL}}(\cdot)$ in Eq. (16) using the CNN, dubbed as VNet as follows:

$$\begin{aligned}\mathbf{V}^{(k)} &= \mathrm{prox}_{\Omega_{NL}}(\mathbf{V}^{(k-1)}, \mathbf{H}^{(k)}) \\ &\approx \mathrm{VNet}(\mathbf{V}^{(k-1)}, \mathbf{N}^{(k)}, \mathbf{P}),\end{aligned} \tag{24}$$

where VNet has similar structure with UNet and its implementation is illustrated in Fig. 4b.

To implement Eqs. (20) and (21) using the network, we firstly split Eq. (21) into three steps:

$$\hat{\mathbf{L}}^{(k)} = \mathbf{D}\mathbf{K}\mathbf{H}^{(k)} \tag{25}$$

$$\mathbf{E}^{(k)} = (\mathbf{D}\mathbf{K})^T(\hat{\mathbf{L}}^{(k)} - \mathbf{L}) \tag{26}$$

$$\mathbf{R}^{(k)} = \mathbf{E}^{(k)} + \eta_1(\mathbf{H}^{(k)} - \mathbf{U}^{(k)}) + \lambda_1(\mathbf{H}^{(k)} - \mathbf{V}^{(k)}). \tag{27}$$

These steps can be transformed into a network, dubbed as HNet, containing many modules corresponding to each operation in the three steps. To be specific, given the $k$-iteration approximated HR target image $\mathbf{H}^{(k)}$, Eq. (25) generates an immediate LR version $\hat{\mathbf{L}}^{(k)}$ by implementing the downsampling $\mathbf{D}$ and low-passing filtering $\mathbf{K}$ functions. This module is defined as

$$\hat{\mathbf{L}}^{(k)} = Down\downarrow_s(\mathbf{H}^{(k)}) \tag{28}$$

where $Down\downarrow_s$ denotes the CNN with the spatial resolution reduction by $s$ times and the reduction is conducted by convolution operator with the $s$ strides.

Followed by, Eq. (26) first computes the LR residual between the input LR target image $\mathbf{L}$ and the generated one $\hat{\mathbf{L}}^{(k)}$, and then acquires the HR residuals $\mathbf{E}^{(k)}$ by applying the corresponding transpose operation $(\mathbf{DK})^T$ to the LR residual. In detail, the transpose function is implemented by the transposed convolutions with $s$ strides as follows:

$$\hat{\mathbf{E}}^{(k)} = Up\uparrow_s(\hat{\mathbf{L}}^{(k)}, \mathbf{L}). \tag{29}$$

Finally, by combining the Eqs. (27) and (20), it is trivial to generate the $k+1$-iteration HR target image $\mathbf{H}^{(k+1)}$ in context of $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$ as follows:

$$\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} - \delta_3\mathbf{R}^{(k)}. \tag{30}$$

To this end, the obtained $\mathbf{U}^{(k)}$, $\mathbf{V}^{(k)}$, and $\mathbf{H}^{(k+1)}$ are transmitted into the next stage.

In summary, the proposed algorithm can be approximated by the unfolding network shown in Fig. 2. The signal flow of the deep unfolding network is

$$\begin{aligned}&\mathbf{H}^{(k)} \Rightarrow \mathbf{U}^{(k)}, \\ &\mathbf{H}^{(k)} \Rightarrow \mathbf{N}^{(k)} \Rightarrow \mathbf{V}^{(k)}, \\ &[\mathbf{U}^{(k)}, \mathbf{V}^{(k)}] \Rightarrow \mathbf{H}^{(k+1)}.\end{aligned} \tag{31}$$

In each stage, the network achieves the sequential updates of the auxiliary variable $\mathbf{U}$, the auxiliary variable $\mathbf{N}$, the auxiliary variable $\mathbf{V}$ and the approximated HR target $\mathbf{H}$, as illustrated in Eq. (31). In each stage, since each module corresponds to the operation in one iteration step, the interpretability of the deep model can thus be improved.

### 3.4 Memory-Augmented Deep Unfolding Network

Nevertheless, there exist several issues of deep unfolding network to be solved. First, the potential of cross-stages, which can be regarded as short-term memory has not been fully explored. In addition, the severe information loss between adjacent stages, recognized as the rarely realized long-term dependency has not been studied due to the feature transformation with channel number reduction, further limiting their improvements.

In this paper, to facilitate the signal flow across iterative stages, the persistent memory mechanism is introduced to augment the information representation by exploiting

the long-short term unit in the image and feature spaces. Specifically, based on the proposed unfolding network in the previous section, we further embed the memory mechanism into it. In the three sub-networks (i.e., UNet, VNet, HNet), the different-layers feature maps and the output intermediate images of each iterative module are selected and integrated for further transformation and then inserted into the next iterative stage for information interaction across stages, thus reducing the information loss. In this way, the information representation can be well improved, leading to better performance. Equipped with above persistent memory mechanism, we propose a memory-augmented deep unfolding network for GISR as shown in Fig. 2. Next, we will introduce the improved version of UNet, VNet, and HNet with embedded memory mechanism in detail.

#### 3.4.1 UNet

To increase the model capability, the memory of previous information at previous stages is introduced to the expressed module corresponding to Eq. (22). As shown in Fig. 4, the UNet is designed with the basic CRB unit which consists of the pure convolution layer and the effective residual blocks. Taking the $k$-th iteration for example, the computation flow of UNet is defined as

$$\mathbf{P}_1^{(k-1)} = \text{Cat}(\text{CRB}(\mathbf{P}), \mathbf{P}) \tag{32}$$

$$\mathbf{H}_1^{(k-1)} = \text{Cat}(\text{CRB}(\mathbf{H}^{(k-1)}), \mathbf{U}_m^{(k-1)}) \tag{33}$$

$$\mathbf{HP}_1^{(k-1)} = \text{Cat}(\mathbf{H}_1^{(k-1)}, \mathbf{P}_1^{(k-1)}) \tag{34}$$

$$\mathbf{HP}_2^{(k-1)} = \text{CRB}(\mathbf{HP}_1^{(k-1)}) \tag{35}$$

$$\mathbf{U}^{(k)} = \text{CRB}(\mathbf{HP}_2^{(k-1)}), \tag{36}$$

where Cat represents the concatenation operation along the channel dimension and $\mathbf{U}_m^{(k-1)}$ is the high-throughput information from previous stage to reduce the information loss. The updated memory $\mathbf{U}^{(k)}$ can be obtained by exploiting ConvLSTM unit to transform the different-layer's features $\mathbf{HP}_1^{(k-1)}$, $\mathbf{HP}_2^{(k-1)}$ and $\mathbf{U}^{(k)}$ as

$$\mathbf{HPU} = \text{CRB}(\text{Cat}(\mathbf{HP}_1^{(k-1)}, \mathbf{HP}_2^{(k-1)}, \mathbf{U}^{(k)})) \tag{37}$$

$$\mathbf{h}_\mathbf{U}^{(k)}, \mathbf{c}_\mathbf{U}^{(k)} = \text{ConvLSTM}(\mathbf{HPU}, \mathbf{h}_\mathbf{U}^{(k-1)}, \mathbf{c}_\mathbf{U}^{(k-1)}) \tag{38}$$

$$\mathbf{U}_m^{(k)} = \text{CRB}(\mathbf{h}_\mathbf{U}^{(k)}) \tag{39}$$

where $\mathbf{h}_\mathbf{U}^{(k-1)}$ and $\mathbf{c}_\mathbf{U}^{(k-1)}$ denotes the hidden state and cell state in ConvLSTM to augment the long-range cross stage information dependency. Furthermore, $\mathbf{h}_\mathbf{U}^{(k)}$ is directly fed into the CRB to generate the updated memory $\mathbf{U}_m^{(k)}$. The transition process of ConvLSTM is unfolded as

$$\mathbf{i}^{(k)} = \sigma(\mathbf{W}_{si} * \mathbf{HPU} + \mathbf{W}_{hi} * \mathbf{h}_\mathbf{U}^{(k-1)} + \mathbf{b}_i), \tag{40}$$

$$\mathbf{f}^{(k)} = \sigma(\mathbf{W}_{sf} * \mathbf{HPU} + \mathbf{W}_{hf} * \mathbf{h}_\mathbf{U}^{(k-1)} + \mathbf{b}_f), \tag{41}$$

$$\mathbf{c}^{(k)} = \mathbf{f}^{(k)} \odot \mathbf{c}_\mathbf{U}^{(k-1)} + \mathbf{i}^{(k)} \odot tanh(\mathbf{W}_{sc} * \mathbf{HPU} + \mathbf{W}_{hc} * \mathbf{h}_\mathbf{U}^{(k-1)} + \mathbf{b}_c), \tag{42}$$

$$\mathbf{o}^{(k)} = \sigma(\mathbf{W}_{so} * \mathbf{HPU} + \mathbf{W}_{ho} * \mathbf{h}_\mathbf{U}^{(k-1)} + \mathbf{b}_o), \tag{43}$$

$$\mathbf{h}_\mathbf{U}^{(k)} = \mathbf{o}^{(k)} \odot tanh(\mathbf{c}_\mathbf{U}^{(k)}) \tag{44}$$

where $*$ and $\odot$ denote the convolution operation and Hadamard product, respectively. $\mathbf{c}_\mathbf{U}^{(k)}$ and $\mathbf{h}_\mathbf{U}^{(k)}$ represent the cell state and hidden state, respectively. $\sigma$ and $tanh$ denote the sigmoid and tanh function, respectively. In this way, not only the information loss of feature channel reduction is alleviated, but also the long-term cross-stage information dependency can be enhanced.

#### 3.4.2 VNet

In Eq. (23), it aims to measure the non-local cross-modality similarity and then aggregates the semantically-related and structure-consistent content from long-range patches in target images and across the modalities of target and guidance images, derived from (Dong et al., 2013). To this end, we devise a novel cross-modality non-local operation module (denoted as CNL). Figure 5 illustrates the CNL module, which receives the updated HR target image $\mathbf{H}^{(k)}$ and guidance image $\mathbf{P}$ as input and generates the refined image $\mathbf{N}^{(k)}$.

Specifically, the updated HR target image feature map $\mathbf{H}^{(k)}$ with the size of $W \times H \times C$ and guidance image feature map $\mathbf{P}$ with the size of $W \times H \times C^{'}$ are transmitted into the CNL. Firstly, we employ two independent $3 \times 3$ convolutions over the target and guidance features, thus reducing the input feature dimension of $\mathbf{H}^{(k)}$ and $\mathbf{P}$ to $\frac{W}{2} \times \frac{H}{2} \times C$, which is the dimension of $\mathbf{H}_r$ and $\mathbf{P}_r$. Secondly, to model the long-range correlation, both the cross-modality and inter-modality operations are conducted. In the left part of Fig. 5, the inter-modality computing imitates the non-local mean filtering over target image feature maps as follows:

$$\mathbf{H}_{r1} = \delta(\mathbf{H}_r), \tag{45}$$

$$\mathbf{H}_{r2} = \theta(\mathbf{H}_r), \tag{46}$$

$$\mathbf{F}_{HH} = \text{softmax}(\mathbf{H}_{r1}\mathbf{H}_{r2}), \tag{47}$$

where the target features $\mathbf{H}_r$ are further processed by the $\delta$ and $\theta$ convolution modules separately, which contain the sequential operations of $3 \times 3$ convolution, the reshape and the near interpolation up-sampling operations, thus generating the two features $\mathbf{H}_{r1}$ and $\mathbf{H}_{r2}$ with size of $\frac{C}{2} \times \frac{WH}{4}$ and $\frac{WH}{4} \times \frac{C}{2}$. Then, the multiplication of $\mathbf{H}_{r1}$ and $\mathbf{H}_{r2}$ is input to a softmax function to generate the attention map $\mathbf{F}_{HH}$.

To exploit the cross-modality correlation, guidance image feature $\mathbf{P}_r$ is passed into the $\phi$ convolution module to obtain

the feature $\mathbf{P}_{r1}$ as follows:

$$\mathbf{P}_{r1} = \phi(\mathbf{P}_r), \tag{48}$$

Then, the cross-modality correlation can be modeled as

$$\mathbf{F}_{HP} = \text{softmax}(\mathbf{H}_{r2}\mathbf{P}_{r1}). \tag{49}$$

Finally, the additional $\varphi(.)$ and $\vartheta(.)$ modules are adopted over $\mathbf{H}^{(k)}$ and $\mathbf{P}$ to provide the the embedding representation $\mathbf{H}_e$ and $\mathbf{P}_e$ with the same size $\frac{WH}{4} \times \frac{C}{2}$. Incorporating the inter-modality $\mathbf{F}_{HH}$ and cross-modality $\mathbf{F}_{HP}$ correlation, the refined feature map $\mathbf{N}^{(k)}$ can be formulated as

$$\mathbf{N}^{(k)} = \text{CRB}(\text{Cat}(\text{C}_1(\mathbf{F}_{HP}\mathbf{P}_e), \text{C}_1(\mathbf{F}_{HH}\mathbf{H}_e))), \tag{50}$$

where $\text{C}_1(.)$ is the the convolution layer with $1 \times 1$ kernel size. With the proposed CNL, it is capable of searching the similarities between long-range patches in target images and across the modalities of target and guidance images, benefiting the texture enhancement.

With the output of CNL module $\mathbf{N}^k$, the previous output $\mathbf{V}^{(k-1)}$ and the accumulated memory state $\mathbf{V}_m^{(k-1)}$, we can obtain the updated $\mathbf{V}^{(k)}$ as shown in Fig. 4b. It can be clearly seen that the VNet has a similar architecture with that of UNet, which is consistent with their similar updating rules. Additionally, the memory transmission of VNet is also the same as that of UNet.

#### 3.4.3 HNet

To transform the update process of $\mathbf{H}^{(k+1)}$, i.e., Eqs. (27), (28), and (29) into a network, firstly, we need to implement the two operations, i.e., $Down \downarrow_s$ and $Up \uparrow_s$, using the network. Specifically, $Down \downarrow_s$ is implemented by a CRB module with spatial identify transformation, and an additional $s$-strides followed CRB module with spatial resolution reduction:

$$\mathbf{KH}^{(k)} = \text{CRB}(\text{Cat}(\mathbf{H}^{(k)}, \mathbf{H}_m^{(k)})) \tag{51}$$

$$\mathbf{DKH}^{(k)} = \text{CRB}^{(s)} \downarrow (\mathbf{KH}^{(k)}) \tag{52}$$

where $\text{CRB}^{(s)} \downarrow$ aims to perform the $s$ times down-sampling. The latter operation $Up \uparrow_s$ is implemented by a transposed convolution containing the $s$-strides CRB module with spatial resolution expansion and a CRB module with spatial identify transformation:

$$\mathbf{UH}^{(k)} = \text{CRB}^{(s)} \uparrow (\mathbf{L} - \mathbf{DKH}^{(k)}) \tag{53}$$

where $\text{CRB}^{(s)} \uparrow$ aims to perform the $s$ times up-sampling. Further, in context of Eqs. (30), (52) and (53), the updated $\mathbf{H}^{(k+1)}$ and the updated memory $\mathbf{H}_m^{(k+1)}$ can be obtained as follows:

$$\mathbf{MH}^{(k+1)} = \text{CRB}(\text{Cat}(\mathbf{KH}^{(k)}, \mathbf{UH}^{(k)}, \mathbf{H}^{(k+1)})) \tag{54}$$

$$\mathbf{h}_{\mathbf{H}}^{(k+1)}, \mathbf{c}_{\mathbf{H}}^{(k+1)} = \text{ConvLSTM}(\mathbf{MH}^{(k+1)}, \mathbf{h}_{\mathbf{H}}^{(k)}, \mathbf{c}_{\mathbf{H}}^{(k)}) \tag{55}$$

$$\mathbf{H}_m^{(k+1)} = \text{CRB}(\mathbf{h}_{\mathbf{H}}^{(k+1)}) \tag{56}$$

where ConvLSTM performs similar functions as aforementioned. The features $\mathbf{KH}^{(k)}$, $\mathbf{UH}^{(k)}$ and $\mathbf{H}^{(k+1)}$ are obtained by different locations, thus possessing more adequate information and alleviate the information loss. Finally, with the output of CNL module $\mathbf{N}^{(k)}$, the updated $\mathbf{V}^{(k)}$, $\mathbf{U}^{(k)}$ and the accumulated memory state $\mathbf{H}_m^{(k)}$, we can obtain the updated $\mathbf{H}^{(k+1)}$ as illustrated in Fig. 4c.

In summary, the signal flow of the proposed memory-augmented deep unfolding network (MADUNet) is

$$\begin{array}{ccccccccccc}
\mathbf{H}_m^{(k-1)} & & \mathbf{U}_m^{(k-1)} & & \mathbf{H}_m^{(k)} & & \mathbf{V}_m^{(k-1)} & & & & \mathbf{H}_m^{(k-1)} \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow & & & & \Downarrow \\
\mathbf{H}^{(k)} & \Rightarrow & \mathbf{U}^{(k)} & \Rightarrow & \mathbf{H}^{(k+1)} & \Leftarrow & \mathbf{V}^{(k)} & \Leftarrow & \mathbf{N}^{(k)} & \Leftarrow & \mathbf{H}^{(k)} \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow & & & & \Downarrow \\
\mathbf{H}_m^{(k)} & & \mathbf{U}_m^{(k)} & & \mathbf{H}_m^{(k+1)} & & \mathbf{V}_m^{(k)} & & & & \mathbf{H}_m^{(k)}
\end{array}$$

### 3.5 Network Training

The training loss for each training pair is defined as the distance between the estimated HR target image from our proposed method and the ground truth HR target image. The most widely used loss function to compute the distance is mean squared error (MSE) loss. However, MSE loss usually generates over-smoothed results. Therefore, we adopt the mean absolute error (MAE) loss to construct our training objective function, which is defined as

$$\mathcal{L} = \sum_{i=1}^{N} \left\| \mathbf{H}_i^{(K+1)} - \mathbf{H}_{gt,i} \right\|_1, \tag{57}$$

where $N$ is the number of training pairs, $\mathbf{H}_i^{(K+1)}$ denote the $i$-th estimated HR target image, and $\mathbf{H}_{gt,i}$ is the $i$-th ground truth HR target image.

## 4 Experiments

In this section, we evaluate the effectiveness and superiority of our proposed method on three typical GISR tasks, i.e., Pan-sharpening (Sect. 4.1), Depth Image SR (Sect. 4.2) and MR Image SR (Sect. 4.3) tasks. In each section, the specific experimental settings and datasets are firstly described. Then, the quantitative and qualitative experimental results

are reported compared with the widely-recognized state-of-the-art methods. Additionally, we conduct an ablation study to gain insight into the respective contributions of the devised components.

## 4.1 Pan-sharpening

To verify the effectiveness of our proposed method on the Pan-sharpening task, we conduct several experiments on the benchmark datasets compared with several representative pan-sharpening methods: (1) Seven commonly-recognized state-of-the-art deep-learning based methods, including PNN (Masi et al., 2016), PANNET (Yang et al., 2017), multiscale and multidepth network (MSDCNN) (Yuan et al., 2018), super-resolution-guided progressive network (SRPPNN) (Cai & Huang, 2020), deep gradient projection network (GPPNN) (Xu et al., 2021), pan-sharpening transformer network (INNformer) (Zhou et al., 2022a) and mutual information-driven pan-sharpening method (MutInf) (Zhou et al., 2022c); (2) Five promising traditional methods, including smoothing filter-based intensity modulation (SFIM) (Liu, 2000), Brovey (Gillespie et al., 1987), GS (Laben & Brower, 2000), intensity hue-saturation fusion (IHS) (Haydn et al., 1982), and PCA guided filter (GFPCA) (Liao et al., 2017).

### 4.1.1 Datasets and Evaluation Metrics

Due to the unavailability of ground-truth pan-sharpened images, we employ the Wald protocol tool (Wald et al., 1997) to generate the training set. Specifically, given the MS image $\mathbf{H} \in R^{M\times N\times C}$ and the PAN image $\mathbf{P} \in R^{rM\times rN}$, both of them are down-sampled with ratio $r$, and then are denoted by $\mathbf{L} \in R^{M/r\times N/r\times C}$ and $\mathbf{p} \in R^{M\times N}$, respectively. Then, $\mathbf{L}$ is regarded as the LR MS image, $\mathbf{p}$ is the guided PAN image, and $\mathbf{H}$ is the ground truth HR MS image. In this experiment, the remote sensing images from three satellites, i.e., WorldView-II, WorldView-III, and GaoFen2, are used for evaluation. Each database contains hundreds of image pairs, and they are divided into training, validation and testing set by 7 : 2 : 1. The number of the training and testing dataset for each satellite is shown in Table 1. Specifically, we employ Wald's protocol (Cao et al., 2021) to generate the training and testing datasets from three satellites, i.e., GaoFen-2, WorldView II, and WorldView III. The generation process mainly contains the following steps: (1) Downsampling the PAN and the MS images by a resolution factor 4 using modulation transfer function (MTF) based filters, thus the downsampled PAN image and the downsampled MS image can be regarded as the training PAN image and the training MS image, respectively; (2) The original MS image is regarded as the training GT image. In the training set, each training pair contains one PAN image with the size of 128 × 128, one LR MS patch with the size of 32 × 32 × 4, and one ground truth HR MS patch with the size of 128 × 128 × 4. For numerical stability, each patch is normalized to [0, 1].

**Table 1** The number of training and testing dataset for each satellite

| Satellite | Gaofen−2 | WorldView II | WorldView III |
|---|---|---|---|
| Training | 4068 | 1140 | 3228 |
| Testing | 400 | 160 | 400 |

To assess the performance of all the methods on the test data with the ground truth, we use the following image quality assessment (IQA) metrics: the relative dimensionless global error in synthesis (ERGAS), the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM), the correlation coefficient (SCC), and the Q index (Vivone et al., 2014).

To evaluate the generalization ability of our method, we create an additional real-world full-resolution dataset of 200 samples over the newly-selected GaoFen2 satellite for evaluation. To be specific, the additional dataset is generated by the full-resolution setting where the PAN and MS images are generated as aforementioned manner without performing the down-sampling, thus PAN image is with the size of 32 × 32 and the MS image with the size of 128 × 128 × 4. Due to the unavailability of ground-truth MS image, we adopt three commonly-used IQA metrics for assessment, i.e., the spectral distortion index $D_\lambda$, the spatial distortion index $D_S$, and the quality without reference (QNR).

### 4.1.2 Implementation Details

In our experiments, all our designed networks are implemented in PyTorch (Paszke et al., 2019) framework and trained on the PC with a single NVIDIA GeForce GTX 3060Ti GPU. In the training phase, these networks are optimized by the Adam optimizer (Kingma & Ba, 2017) over 1000 epochs with a mini-batch size of 4. The learning rate is initialized with $8 \times 10^{-4}$. When reaching 200 epochs, the learning rate is decayed by multiplying 0.5. Furthermore, all the hidden and cell states of ConvLSTM are initialized as zero and the input $\mathbf{H}^{(0)}$ of our unfolding network is obtained by applying Bibubic up-sampling over LR target image $\mathbf{L}$.

### 4.1.3 Comparison with SOTA Methods

In this section, we will perform the detailed quantitative and qualitative experimental analysis over datasets from WorldView-III, WorldView-II and GaoFen2 satellites to demonstrate the effectiveness of our proposed method.

*WorldView-III Dataset Results* We provide a qualitative assessment of the selected competitive techniques for the WordView-III dataset. The average metric results in terms of

**Table 2** The average quantitative results on the WorldView-III dataset

| Methods | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ | SCC ↑ | Q ↑ |
|---|---|---|---|---|---|---|
| SFIM (Liu, 2000) | 21.8212 | 0.5457 | 0.1208 | 8.973 | 0.6952 | 0.4531 |
| GS (Laben & Brower, 2000) | 22.5608 | 0.547 | 0.1217 | 8.2433 | 0.7131 | 0.4411 |
| Brovey (Gillespie et al., 1987) | 22.506 | 0.5466 | 0.1159 | 8.2331 | 0.7033 | 0.4394 |
| IHS (Haydn et al., 1982) | 22.5579 | 0.5354 | 0.1266 | 8.3616 | 0.6994 | 0.4301 |
| GFPCA (Liao et al., 2017) | 22.34 | 0.4826 | 0.1294 | 8.3964 | 0.6987 | 0.3115 |
| PNN (Masi et al., 2016) | 29.9418 | 0.9121 | 0.0824 | 3.3206 | 0.954 | 0.8679 |
| PANNet(Yang et al., 2017) | 29.684 | 0.9072 | 0.0851 | 3.4263 | 0.9512 | 0.8631 |
| MSDCNN(Yuan et al., 2018) | 30.3038 | 0.9184 | 0.0782 | 3.1884 | 0.9577 | 0.8763 |
| SRPPNN (Cai & Huang, 2020) | 30.4346 | 0.9202 | 0.0770 | 3.1553 | 0.9581 | 0.8776 |
| GPPNN (Xu et al., 2021) | 30.1785 | 0.9175 | 0.0776 | 3.2593 | 0.9569 | 0.8739 |
| MutInf (Zhou et al., 2022c) | 30.4907 | 0.9223 | 0.0749 | 3.1125 | 0.9569 | 0.8802 |
| INNformer (Zhou et al., 2022a) | 30.5365 | 0.9225 | 0.0747 | 3.0997 | 0.9569 | 0.8817 |
| Ours | **30.5451** | **0.9214** | **0.0769** | **3.1032** | **0.9598** | **0.8804** |

The best performance are shown in bold

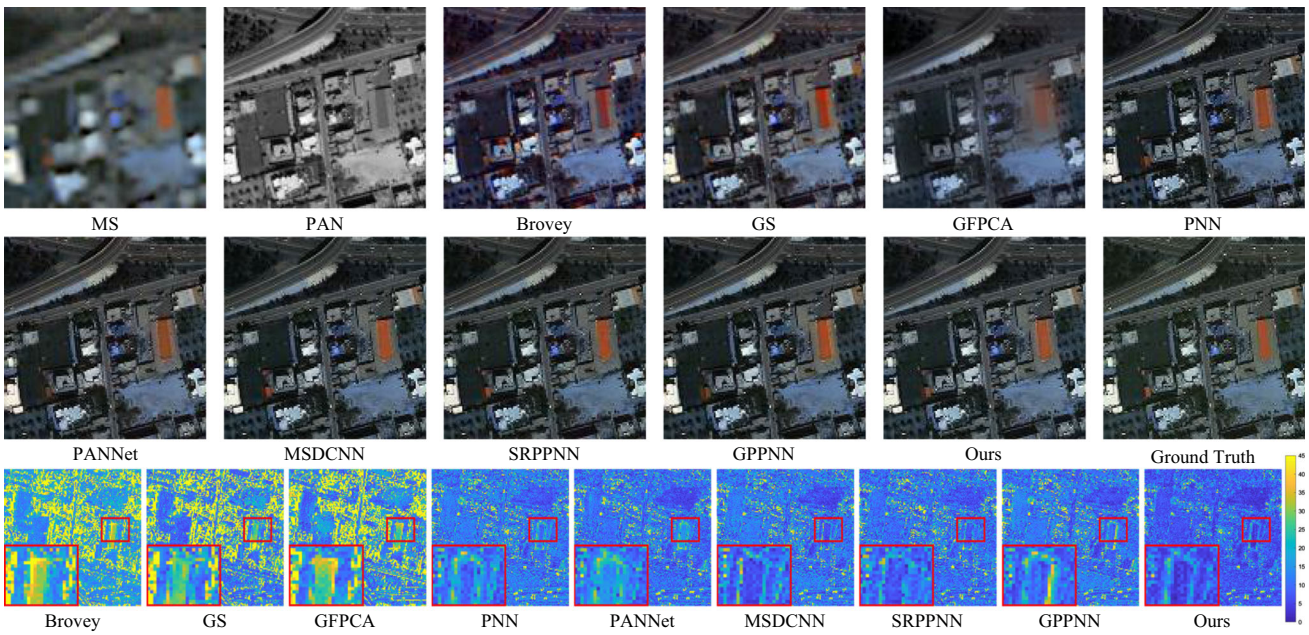



**Fig. 6** Visual comparisons of the fused HRMS image for all the methods on one WorldView-III dataset. Images in the last row visualize the MSE between the pan-sharpened results and the ground truth

all adopted IQAs for all competing methodologies are tabulated in Table 2. It is self-evident that our proposed method outperforms all other competing methods across all metric indexes. Deepening into the comparison of the competing deep learning methods, our method outperforms the pioneer Pan-sharpening network PNN (Masi et al., 2016) by 0.6033 dB and 0.0093 in PSNR and SSIM, and by 0.0055, 0.2174 in SAM and ERGAS, suggesting the improved spatial information augmentation and reduced spectrum distortion. Especially for SRPPNN with the most parameters (Cai & Huang, 2020), our method outperforms it by 0.1105 dB, 0.0012 in PSNR and SSIM, and is lower by 0.0001, 0.0521 in SAM and ERGAS, while using less parameters. In detail, our model requires only 0.8 M while SRPPNN requires 17 M, illustrating the efficacy of our method even more. Considering the remaining competing methodologies, identical evidence may be detected, and the consistent result can be supported by the WorldView-III dataset. In terms of visual comparison, Fig. 6 displays all the generated images and the MSE residual map for the most promising comparative approaches on the pseudocolor map in the last row. As can be observed from these results, our proposed method achieves the best performance than all the other compared methods

**Table 3** The average quantitative results on the WorldView-II dataset

| Methods | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ | SCC ↑ | Q ↑ |
|---|---|---|---|---|---|---|
| SFIM (Liu, 2000) | 34.1297 | 0.8975 | 0.0439 | 2.3449 | 0.9079 | 0.6064 |
| GS (Laben & Brower, 2000) | 35.6376 | 0.9176 | 0.0423 | 1.8774 | 0.9225 | 0.6307 |
| Brovey (Gillespie et al., 1987) | 35.8646 | 0.9216 | 0.0403 | 1.8238 | 0.8913 | 0.6163 |
| IHS (Haydn et al., 1982) | 35.2962 | 0.9027 | 0.0461 | 2.0278 | 0.8534 | 0.5704 |
| GFPCA (Liao et al., 2017) | 34.558 | 0.9038 | 0.0488 | 2.1401 | 0.8924 | 0.4665 |
| PNN (Masi et al., 2016) | 40.7550 | 0.9624 | 0.0259 | 1.0646 | 0.9677 | 0.7426 |
| PANNet (Yang et al., 2017) | 40.8176 | 0.9626 | 0.0257 | 1.0557 | 0.968 | 0.7437 |
| MSDCNN (Yuan et al., 2018) | 41.3355 | 0.9664 | 0.0242 | 0.9940 | 0.9721 | 0.7577 |
| SRPPNN (Cai & Huang, 2020) | 41.4538 | 0.9679 | 0.0233 | 0.9899 | 0.9729 | 0.7691 |
| GPPNN (Xu et al., 2021) | 41.1622 | 0.9684 | 0.0244 | 1.0315 | 0.9722 | 0.7627 |
| MutInf (Zhou et al., 2022c) | 41.6773 | 0.9705 | 0.0224 | 0.9519 | 0.9751 | 0.7753 |
| INNformer (Zhou et al., 2022a) | 41.6903 | 0.9704 | 0.0227 | 0.9514 | 0.9753 | 0.7756 |
| Ours | **41.8577** | **0.9697** | **0.0229** | **0.9420** | **0.9745** | **0.7740** |

The best performance are shown in bold

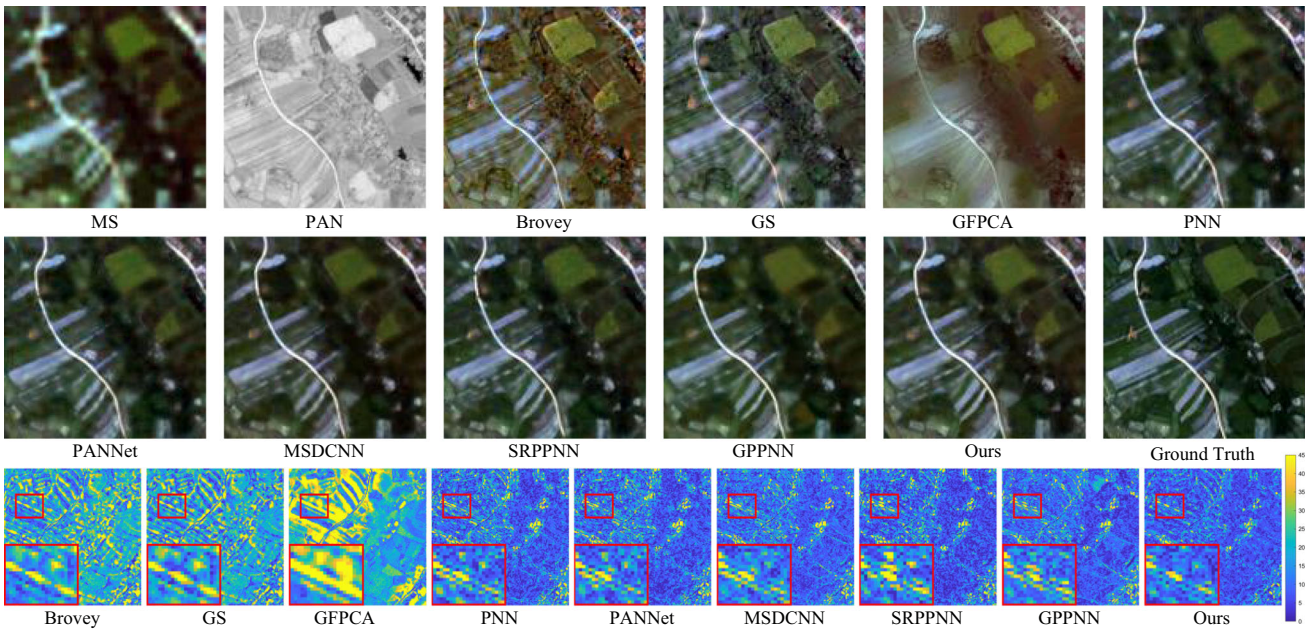



**Fig. 7** Visual comparisons of the fused HR MS image for all the methods on one WorldView-II dataset. Images in the last row visualizes the MSE between the pan-sharpened results and the ground truth

by accurately enhancing the spatial details and maintaining the spectral information.

*WorldView-II Dataset Results* Table 3 presents the average quantitative performance between our method and aforementioned competitive algorithms over the WorldView-II dataset. From Table 3, it can be observed that our proposed method can significantly outperform other state-of-the-art competing methods in terms of all the metrics. To summarize at a high level, it is clearly figured out that deep learning based methods surpass the traditional methods, attributing to the powerful learning capability of deep neural networks. In particular, compared with PNN, our method is higher by 1.1027 dB and 0.0073 in PSNR and SSIM, and lower by 0.003, 0.1226 in SAM, and ERGAS, indicating better spatial information enhancement and lower spectral distortion. Compared with SRPPNN (Cai & Huang, 2020), our method performs better than SRPPNN with a large margin by 0.4039 dB on PSNR and 0.0018 on SSIM. As for the remaining metrics, our method also achieves better results than SRPPNN. Deepening into the definition of all the metrics, the best performance of our method demonstrates that our proposed method is capable of preserving precise spatial structures and avoiding the spectral distortion. In addition, we also present the visual results comparison of all the methods on one rep-

**Table 4** The average quantitative results on the GaoFen2 dataset

| Methods | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ | SCC ↑ | Q ↑ |
|---|---|---|---|---|---|---|
| SFIM (Liu, 2000) | 36.906 | 0.8882 | 0.0318 | 1.7398 | 0.8128 | 0.4349 |
| GS (Laben & Brower, 2000) | 37.226 | 0.9034 | 0.0309 | 1.6736 | 0.7851 | 0.4211 |
| Brovey (Gillespie et al., 1987) | 37.7974 | 0.9026 | 0.0218 | 1.3720 | 0.6446 | 0.3857 |
| IHS (Haydn et al., 1982) | 38.1754 | 0.91 | 0.0243 | 1.5336 | 0.6738 | 0.3682 |
| GFPCA (Liao et al., 2017) | 37.9443 | 0.9204 | 0.0314 | 1.5604 | 0.8032 | 0.3236 |
| PNN (Masi et al., 2016) | 43.1208 | 0.9704 | 0.0172 | 0.8528 | 0.9400 | 0.739 |
| PANNet (Yang et al., 2017) | 43.0659 | 0.9685 | 0.0178 | 0.8577 | 0.9402 | 0.7309 |
| MSDCNN (Yuan et al., 2018) | 45.6874 | 0.9827 | 0.0135 | 0.6389 | 0.9526 | 0.7759 |
| SRPPNN (Cai & Huang, 2020) | 47.1998 | 0.9877 | 0.0106 | 0.5586 | 0.9564 | 0.7900 |
| GPPNN (Xu et al., 2021) | 44.2145 | 0.9815 | 0.0137 | 0.7361 | 0.9510 | 0.7721 |
| MutInf (Zhou et al., 2022c) | 47.3042 | 0.9892 | 0.0102 | 0.5481 | 0.9603 | 0.8025 |
| INNformer (Zhou et al., 2022a) | 47.3528 | 0.9893 | 0.0102 | 0.5479 | 0.9611 | 0.8037 |
| Ours | **47.2668** | **0.9890** | **0.0102** | **0.5472** | **0.9597** | **0.7973** |

The best performance are shown in bold

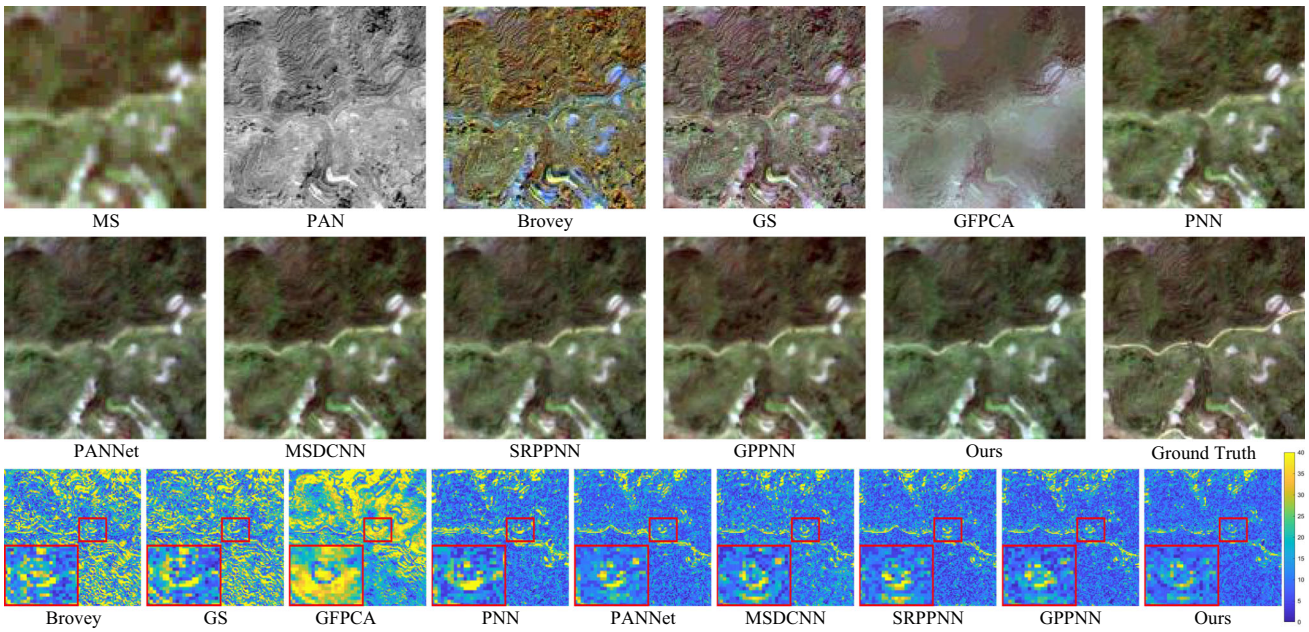



**Fig. 8** Visual comparisons of the fused HRMS image for all the methods on one GaoFen2 dataset. Images in the last row visualizes the MSE between the pan-sharpened results and the ground truth

**Table 5** The average quantitative results on the GaoFen2 datasets in the full resolution case

| Metrics | SFIM | GS | Brovey | IHS | GFPCA | PNN | PANNET | MSDCNN | SRPPNN | GPPNN | **Ours** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_\lambda$ ↓ | 0.0822 | 0.0696 | 0.1378 | 0.0770 | 0.0914 | 0.0746 | 0.0737 | 0.0734 | 0.0767 | 0.0782 | **0.0695** |
| $D_s$ ↓ | **0.1087** | 0.2456 | 0.2605 | 0.2985 | 0.1635 | 0.1164 | 0.1224 | 0.1151 | 0.1162 | 0.1253 | 0.1139 |
| QNR ↑ | 0.8214 | 0.7025 | 0.6390 | 0.6485 | 0.7615 | 0.8191 | 0.8143 | 0.8251 | 0.8173 | 0.8073 | **0.8235** |

The best performance are shown in bold

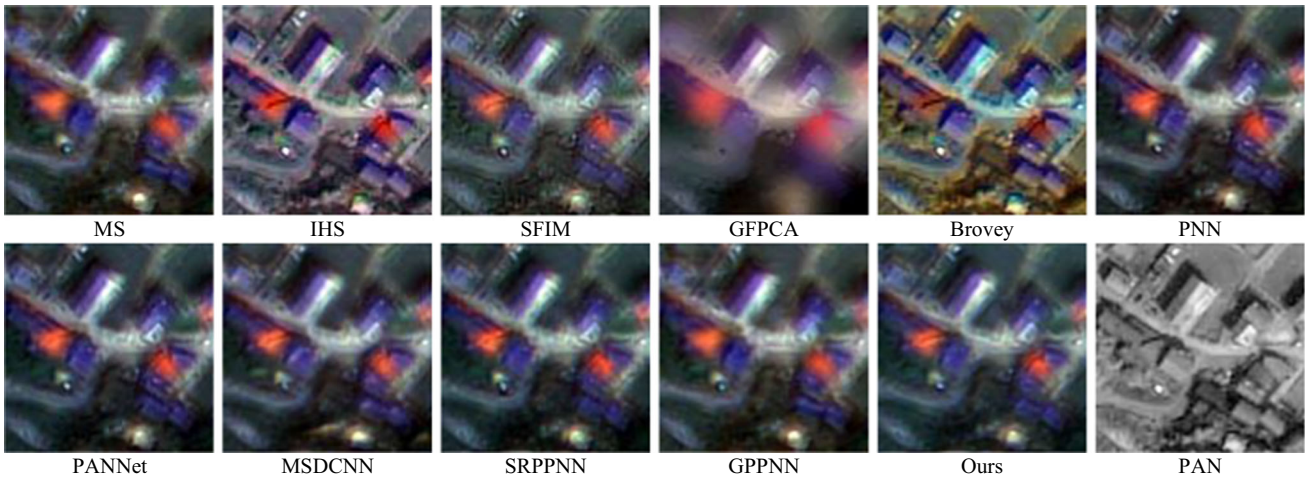



**Fig. 9** Visual comparisons of the fused HRMS image for all the methods on a full resolution sample

resentative sample from WorldView-II dataset in Fig. 7. To highlight the differences in detail, we select the R, G, B bands of the generated MS images to better visualize the qualitative comparison. As can be seen, our method can obtain better visual effect since it accurately enhance the spatial details and preserve the spectral information, which is consistent with quantitative results shown in Table 3. Also, the MSE residual between the generated MS image and the ground truth HR MS image are presented in the last row of Fig. 7, from which we can see that traditional methods, i.e., SFIM, Brovey, and GS, suffer from notable artifacts, while the residual map of our method contains fewer details than other methods, which further verifies the superiority of our network.

*Gaofen2 Dataset Results* Table 4 summarizes the average quantitative results of all the methods on the Gaofen2 dataset. Clearly, our proposed method outperforms other competing methods in terms of all the indexes. We can draw a similar conclusion with the Worldview II dataset. For example, deep-learning-based algorithms performs much better than classical methods. It can be seen from Table 4 that our method surpasses the second best SRPPNN (Cai & Huang, 2020) by 0.067 dB, 0.0013 in PSNR and SSIM, and is lower by 0.0004, 0.0114 in SAM and ERGAS. To make a visual comparison, Fig. 8 presents all generated pan-sharpened images and the MSE residual map between the pan-sharpened images and the ground truth images. As can be seen, our method can generate better visual images due to its better ability of enhancing the spatial details and preserving the spectral information. This visual result is consistent with the quantitative results shown in Table 4. For easy comparison, we also draw the MSE residual map between the generated MS image and the ground truth HR MS image in the last row of Fig. 8, and one small area of the residual map is amplified, from which we can observe that our method contains less structure and detail information, further verifying the superiority of our method.

#### 4.1.4 Effect on Full-resolution Scenes

To assess the performance of our network in the full resolution case and the model generalization ability, we apply a pre-trained model built on GaoFen2 data to some unseen GaoFen2 satellite datasets with PAN of $32 \times 32$ and MS of $128 \times 128 \times 8$ resolution samples. Specifically, the additional datasets are constructed using the full-resolution setting, in which the PAN and MS images are generated in their original scale using the same methods as in the preceding Sect. 4.1.1, but down-sampled. The experimental results of the all the methods are summarized in Table 5. From Table 5, we can observe that our proposed method performs almost the best in terms of all the indexes, which indicates that our method has better generalization ability compared with other traditional and deep learning-based methods. This is due to that our method is a model-driven deep learning method, which embeds the prior knowledge of the data into the network design and thus the proposed network naturally inherits some generalization ability of the model-based methods. Additionally, we also show visual comparisons for all the methods on a full-resolution sample in Fig. 9, from which we can observe that our proposed network obtains better visual fused effect both spatially and spectrally than other competing approaches.

#### 4.1.5 Complexity Analysis

To conduct a more thorough analysis, we investigate the complexity of the proposed method, including the floating-point operations (FLOPs), the number of parameters (in 10 M) and running time in this section. Comparisons on parameter numbers and model performance (as measured by PSNR) are provided in Table 6 and Fig. 10. As can be seen, PNN (Masi et al., 2016) and PANNet (Yang et al., 2017) have the fewest

**Table 6** Comparisons on parameter numbers, FLOPs and time

| Methods | PNN | PANNet | MSDCNN | SRPPNN | GPPNN | Ours |
|---|---|---|---|---|---|---|
| #Params | 0.689 | 0.688 | 2.390 | 17.114 | 1.198 | 0.700 |
| FLOPs | 1.1289 | 1.1275 | 3.9158 | 21.1059 | 1.3967 | 4.4543 |
| Time (s) | 0.0028 | 0.0027 | 0.0037 | 0.0098 | 0.0075 | 0.0081 |

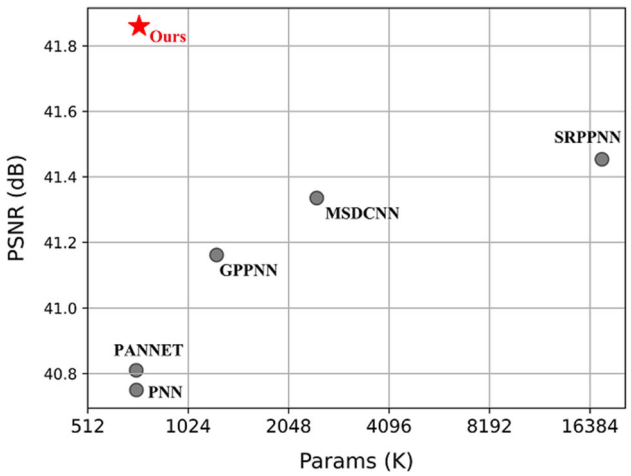



**Fig. 10** Comparisons of model performance and number of parameters

FLOPs due to their network architectures containing only a few convolutional layers. Additionally, they extract a small number of feature maps, resulting in a smaller FLOPs count. MSDCNN (Yuan et al., 2018) and SRPPNN (Cai & Huang, 2020) both exhibit large increases in the number of parameters and FLOPs as the number of convolutional layers and the complexity of the network design rise. Notably, they also achieve the promising performance at the expense of massive model computation and storage. Additionally, the most comparable solution to ours, GPPNN (Xu et al., 2021), is organized around the model-based unfolding principle and has comparable model parameters and flops reductions but inferior performance. This is due to powerful model learning's incapability without fully exploring the potential of different modalities. Also, our method is comparable with other methods in terms of running time. In summary, our network achieves a favorable trade-off between calculation, storage, and model performance compared with other methods.

## 4.2 Depth Image SR

In this section, to verify the effectiveness of our method, we conduct several experiments to compare our method with the representative state-of-the-art Depth Image SR methods. Following the experimental protocol of (Kim et al., 2021), we generate the evaluated datasets over two down-samping operations, i.e., bibubic down-sampling and direct down-sampling. The detailed analysis is described as bellow.

### 4.2.1 Datasets and Metrics

*Dataset* NYU v2 dataset (Silberman et al., 2012) is the widely-recognized benchmark for Depth Image SR. This dataset consists of 1449 RGB-D image pairs recorded by Microsoft Kinect sensors in the structured light. Following the same settings as previous Depth Image SR methods (Li et al., 2019; Kim et al., 2021), we train our proposed network over the first 1000 RGB-D image pairs and then evaluate the trained model on the remaining 449 RGB-D images pairs. We follow the experimental protocol of (Kim et al., 2021) to build the low-resolution depth map, which involves using the bibubic and direct down-sampling operations at different ratios ($\times 4$, $\times 8$ and $\times 16$), respectively. In addition, the root mean squared error (RMSE) is adopted by default to evaluate the model performance.

Furthermore, to determine the potential generalization ability of the model, following the same setting as (Kim et al., 2021), we directly test the trained model over NYU v2 dataset over the additional Middlebury dataset (Scharstein & Pal, 2007) and Lu (Lu et al., 2014) dataset, which are another two benchmark datasets to evaluate the performance of Depth Image SR algorithms. The Middlebury dataset contains 30 RGB-D image pairs, of which 21 pairs are from 2001 (Scharstein & Szeliski, 2002) and 9 pairs are from 2006 (Hirschmuller & Scharstein, 2007). The Lu dataset contains 6 RGB-D image pairs. Following the existing works (Guo et al., 2018; Ye et al., 2020; Kim et al., 2021), we quantify all the recovered depth maps to 8-bits before calculating the MAE or RMSE values for fair evaluation. For both criteria, lower values indicate higher performance.

### 4.2.2 Implementation Details

In the training phase, we choose the released resource code of (Kim et al., 2021) as baseline to implement all our experiments. Specifically, all the our networks are optimized the Adam optimizer (Kingma & Ba, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\mathrm{e}-8$. The initial learning rate is $2 \times 10^{-4}$. For every 100 epochs, the learning rate is decayed by multiplying 0.5. Furthermore, all the hidden and cell states of ConvLSTM are initialized as zero and the input $\mathbf{H}^{(0)}$ of our unfolding network is obtained by applying bibubic upsampling over LR target image $\mathbf{L}$.

### 4.2.3 Experimental Results on Bibubic Down-sampling

To evaluate the effectiveness of the proposed method on the bibubic down-sampling case, following the same setting as (Kim et al., 2021), we select the representative state-

**Table 7** Average RMSE performance comparison for scale factors 4×, 8× and 16× with bicubic down-sampling

| Method | Middlebury | | | Lu | | | NYU v2 | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4× | 8× | 16× | 4× | 8× | 16× | 4× | 8× | 16× | 4× | 8× | 16× |
| Bicubic | 2.47 | 4.65 | 7.49 | 2.63 | 5.23 | 8.77 | 4.71 | 8.29 | 13.17 | 3.27 | 6.06 | 9.81 |
| GF (He et al., 2013) | 3.24 | 4.36 | 6.79 | 4.18 | 5.34 | 8.02 | 5.84 | 7.86 | 12.41 | 4.42 | 5.85 | 9.07 |
| TGV (Ferstl et al., 2013) | 1.87 | 6.23 | 17.01 | 1.98 | 6.71 | 18.31 | 3.64 | 10.97 | 39.74 | 2.50 | 7.97 | 25.02 |
| DGF (Wu et al., 2018b) | 1.94 | 3.36 | 5.81 | 2.45 | 4.42 | 7.26 | 3.21 | 5.92 | 10.45 | 2.53 | 4.57 | 7.84 |
| DJF (Li et al., 2016a) | 1.68 | 3.24 | 5.62 | 1.65 | 3.96 | 6.75 | 2.80 | 5.33 | 9.46 | 2.04 | 4.18 | 7.28 |
| DMSG (Hui et al., 2016) | 1.88 | 3.45 | 6.28 | 2.30 | 4.17 | 7.22 | 3.02 | 5.38 | 9.17 | 2.40 | 4.33 | 7.17 |
| DJFR (Li et al., 2019) | 1.32 | 3.19 | 5.57 | 1.15 | 3.57 | 6.77 | 2.38 | 4.94 | 9.18 | 1.62 | 3.90 | 7.17 |
| DSRNet (Guo et al., 2018) | 1.77 | 3.05 | 4.96 | 1.77 | 3.10 | 5.11 | 3.00 | 5.16 | 8.41 | 2.18 | 3.77 | 6.16 |
| PacNet (Su et al., 2019) | 1.32 | 2.62 | 4.58 | 1.20 | 2.33 | 5.19 | 1.89 | 3.33 | 6.78 | 1.47 | 2.76 | 5.53 |
| FDKN (Kim et al., 2021) | 1.08 | 2.17 | 4.50 | **0.82** | 2.10 | 5.05 | 1.86 | 3.58 | 6.96 | 1.25 | 2.62 | 5.50 |
| DKN (Kim et al., 2021) | 1.23 | 2.12 | 4.24 | 0.96 | 2.16 | 5.11 | 1.62 | 3.26 | 6.51 | 1.27 | 2.51 | 5.29 |
| Ours | **1.15** | **1.69** | **3.23** | 0.90 | **1.74** | **3.86** | **1.51** | **3.02** | **6.23** | **1.18** | **2.15** | **4.44** |

The best performance is shown in bold and second best performance is underlined

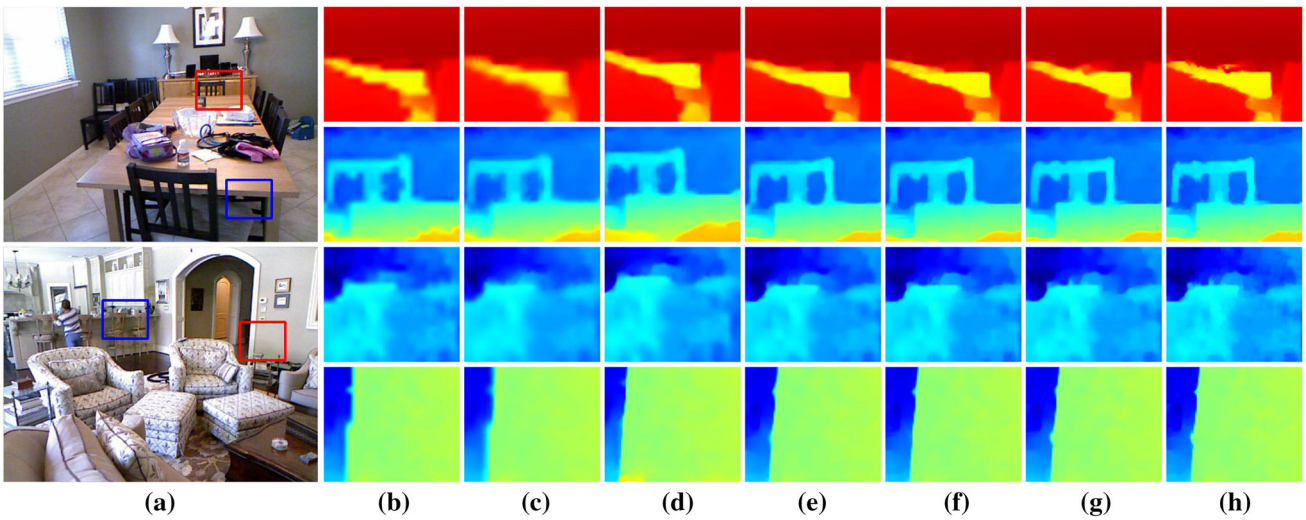



**Fig. 11** Visual comparison of 4× upsampling results: **a** RGB image, **b** Bibubic, **c** GF, **d** DJF, **e** PacNet, **f** DKN [9], **g** Ours, **h** GT

of-the-art Depth Image SR algorithms for comparison: 1) traditional methods, including guided image filtering (GF) He et al. (2013) and total generalized variation (TGV) (Ferstl et al., 2013), 2) deep learning based methods, containing DGF (Wu et al., 2018b), DJF (Li et al., 2016a), DMSG (Hui et al., 2016), deep joint image filte (DJFR) (Li et al., 2019), depth super-resolution network (DSRNet) (Guo et al., 2018), pixel-adaptive convolution (PacNet) (Su et al., 2019), fast deformable Kernel Networks (FDKN) (Kim et al., 2021) and deformable Kernel Networks (DKN) (Kim et al., 2021). All the results are obtained from the original papers reported by the authors for fair comparison. Table 7 reports the quantitative comparison among all the competing algorithms at different ratios, i.e., ×4, ×8 and ×16. The average RMSE values between the generated HR depth map and the ground truth depth map are illustrated.

From Table 7, we can observe that our proposed network performs the best compared with other methods at different ratios of ×4, ×8 and ×16 in average RMSE values. To highlight at a high level, deep-learning based methods (Kim et al., 2021; Guo et al., 2018; Su et al., 2019; Hui et al., 2016; Li et al., 2019) achieve better results than traditional methods (He et al., 2013; Ferstl et al., 2013) by significant margins in terms of RMSE, which is attributed to the powerful learning and mapping capability of deep neural networks. Compared with the second best method DKN (Kim et al., 2021), our method decreases the average RMSE over all the three datasets by 0.09 (4×), 0.36 (8×) and 0.85 (16×), respectively. To evaluate the generalization ability, the models are well trained over NYU v2 dataset and not further fine-tuned over other datasets (i.e., Middlebury dataset and Lu dataset). From Table 7, it can be seen that our method obtains superior performance on the

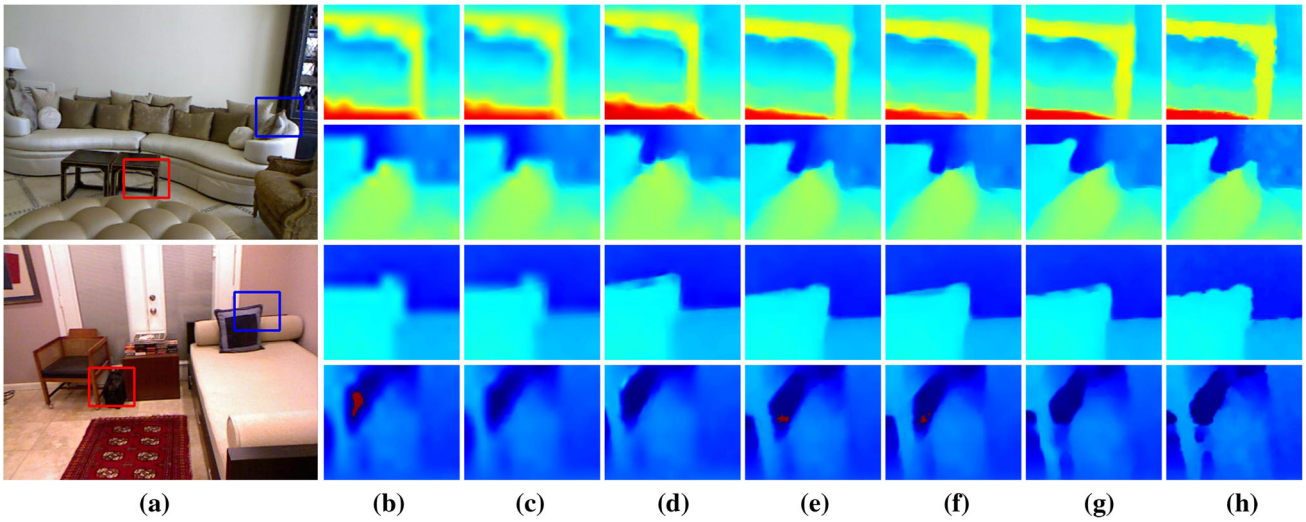



**Fig. 12** Visual comparison of 8× upsampling results: **a** RGB image, **b** Bibubic, **c** GF, **d** DJF, **e** PacNet, **f** DKN, **g** Ours, **h** GT

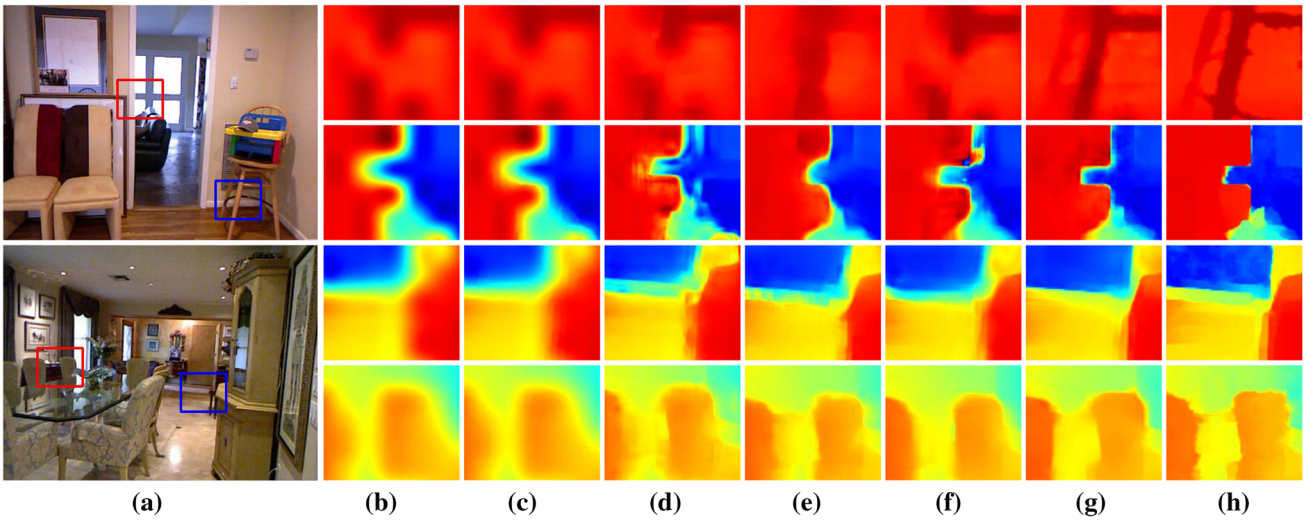



**Fig. 13** Visual comparison of 16× upsampling results: **a**: RGB image, **b** Bibubic, **c** GF, **d** DJF, **e** PacNet, **f** DKN, **g** Ours, **h** GT

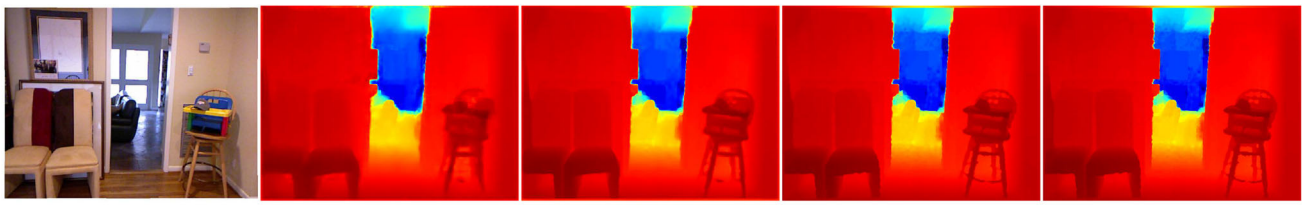
Middlebury dataset and Lu dataset, which implies that our method has better generalization ability compared with other methods.

In terms of visual comparison, Figs. 11, 12 and 13 display all the generated latent HR depth maps. As can be seen, other competing methods exist some issues. For example, GF (He et al., 2013) generates over-smoothed map since the local filter cannot capture global information. DJFR (Li et al., 2019) and DKN (Kim et al., 2021) suffer from diffusion artifacts. PacNet (Su et al., 2019) can preserve the local details, but cannot reconstruct the boundary well. On the contrary, our proposed method obtains the best visual effect than other competing methods since it can enhance the spatial details of LR depth maps, and generate accurate and sharp edges.

Additionally, we also select one representative sample to verify the effect difference over different ratios, as shown in Fig. 14, from which we can see that the ×4 up-sampling results are more reasonable than that of ×8 and ×16. The reason is that the higher ratios of down-sampling loss much more information than the lower ratio case, thus resulting in that the reconstruction process to be difficult. The consistent conclusion can also be supported by the quantitative results reported in Table 7.

#### 4.2.4 Experimental Results on Direct Down-sampling

In terms of the direct down-sampling case, we compare the proposed method with the following state-of-the-art methods: directly Bibubic up-sampling, markov random fields (MRF) (Diebel & Thrun, 2005), guided image filtering (GF) (He et al., 2013), total generalized variation network (TGV) (Ferstl et al., 2013), high quality depth



**(a)** RGB image. **(b)** ×16. **(c)** ×8. **(d)** ×4. **(e)** GT.

**Fig. 14** Visual comparison between 16×, 8× and 4× upsampling results of our method

map up-sampling (Park) (Park et al., 2011), joint static and dynamic guidance (Ham et al., 2015), joint bilateral up-sampling (JBU) (Kopf et al., 2007a), deep joint image filtering (DJF) (Li et al., 2016a), deep multi-scale guidance network (DMSG) (Hui et al., 2016), pixel-adaptive convolution neural network (PacNet) Su et al. (2019), deep joint image filter (DJFR) (Li et al., 2019), depth super-resolution network (DSRNet) (Guo et al., 2018), fast deformable kernel network (DKN) (Kim et al., 2021) and deformable kernel network (DKN) (Kim et al., 2021). All the results for other competing methods are directly taken from the original paper reported by the authors for fair comparison. Table 8 presents the quantitative comparison among all the competing algorithms at different ratios, i.e., 4×, 8× and 16×. The average RMSE values are recorded.

In comparison with other approaches, our method obtains the best performance in terms of the average RMSE values for all the scaling factors. In the most difficult ×16 case for all the datasets, our proposed method obviously outperforms other methods as shown in Table 8. Specifically, compared with the second best method DKN (Kim et al., 2021), our method decreases the average RMSE by 0.15 (4×), 0.23 (8×) and 0.19 (16×), respectively. Additionally, to evaluate the generalization ability, our model is trained over the NYU v2 dataset and then is directly used for testing on the other two datasets, i.e., Middlebury dataset and Lu dataset. From Table 8, we can observe that our method achieves better performance on the Middlebury dataset and Lu dataset compared with other methods, which implies that our method has good generalization ability. In particular, compared with the second best model DKN (Kim et al., 2021), our method decreases the average RMSE by 0.11 (4×), 0.25 (8×) and 0.35 (16×) on the Middlebury dataset (Scharstein & Pal, 2007), and 0.22 (4×), 0.28 (8×) and 0.03 (16×) on the Lu (Lu et al., 2014) dataset.

#### 4.2.5 Parameter Comparison

To conduct a more thorough evaluation, we also investigate the complexity of our method on the parameter number. Table 9 records the parameter number and model performance (evaluated by RMSE index) in the ratio ×4 case. As can be seen, our method obtains the best performance while contains the second fewest parameters. As for other competing methods, DJFR (Li et al., 2019) and PacNet (Su et al., 2019) have the fewest model storage due to their simple network architectures, but results in poor performance. DSRnet (Guo et al., 2018) and PMBAN (Ye et al., 2020) both exhibit large increases in the number of parameters, but they also achieve a slightly decrease on the RSME value. Additionally, DKN (Kim et al., 2021) achieve the second best performance but has more parameters than our method. To emphasize, our network achieves a favorable trade-off between model complexity and model performance compared with other state-of-the-art methods.

### 4.3 MRI Image SR

This section conducts a serious of experiments to evaluates the performance our method on the MR Image SR task. Several representative MR Image SR methods are used for comparison, including Bibubic up-sampling, plain convolutional neural network (PCNN) (Zeng et al., 2018), progressive network (PGN) (Lyu et al., 2020), multi-stage integration network (MINet) (Feng et al., 2021).

#### 4.3.1 Datasets and Implementation Details

In MR Image SR, it produces different contrast images of T1 and T2 but with the same anatomical structure. Due to the complementary property of T1 and T2, MR Image SR aims to super-solve the low-spatial resolution T1 image with the guidance of high-resolution T2. Following the setting of (Zeng et al., 2018), we generates the MR Image SR datasets for two types of settings as bellow. Given the T1 and T2 images with the size of 300 × 256 × 1, T1 is down-sampled into 75 × 64 × 1 with ratio 4, and 150 × 128 × 1 resolutions with ratio 2. The generated dataset is split into the training, validation and testing part by 7 : 2 : 1 respectively, and each one contains hundreds of samples. To assess the model performance, the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) are adopted.

In our experiments, all our designed networks are implemented in PyTorch (Paszke et al., 2019) framework and

**Table 8** Average RMSE performance comparison for scale factors 4×, 8× and 16× with direct down-sampling

| Method | Middlebury | | | Lu | | | NYU v2 | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4× | 8× | 16× | 4× | 8× | 16× | 4× | 8× | 16× | 4× | 8× | 16× |
| Bicubic | 4.44 | 7.58 | 11.87 | 5.07 | 9.22 | 14.27 | 8.16 | 14.22 | 22.32 | 5.89 | 10.34 | 16.15 |
| MRF (Diebel & Thrun, 2005) | 4.26 | 7.43 | 11.80 | 4.90 | 9.03 | 14.19 | 7.84 | 13.98 | 22.20 | 5.67 | 10.15 | 16.06 |
| GF (He et al., 2013) | 4.01 | 7.22 | 11.70 | 4.87 | 8.85 | 14.09 | 7.32 | 13.62 | 22.03 | 5.40 | 9.90 | 15.94 |
| TGV (Ferstl et al., 2013) | 3.39 | 5.41 | 12.03 | 4.48 | 7.58 | 17.46 | 6.98 | 11.23 | 28.13 | 4.95 | 8.07 | 19.21 |
| Park (Park et al., 2011) | 2.82 | 4.08 | 7.26 | 4.09 | 6.19 | 10.14 | 5.21 | 9.56 | 18.10 | 4.04 | 6.61 | 11.83 |
| Ham (Ham et al., 2015) | 3.14 | 5.03 | 8.83 | 4.65 | 7.73 | 11.52 | 5.27 | 12.31 | 19.24 | 4.35 | 8.36 | 13.20 |
| JBU (Kopf et al., 2007a) | 2.44 | 3.81 | 6.13 | 2.99 | 5.06 | 7.51 | 4.07 | 8.29 | 13.35 | 3.17 | 5.72 | 9.00 |
| DGF (Wu et al., 2018b) | 3.92 | 6.04 | 10.02 | 2.73 | 5.98 | 11.73 | 4.50 | 8.98 | 16.77 | 3.72 | 7.00 | 12.84 |
| DJF (Li et al., 2016a) | 2.14 | 3.77 | 6.12 | 2.54 | 4.71 | 7.66 | 3.54 | 6.20 | 10.21 | 2.74 | 4.89 | 8.00 |
| DMSG (Hui et al., 2016) | 2.11 | 3.74 | 6.03 | 2.48 | 4.74 | 7.51 | 3.37 | 6.20 | 10.05 | 2.65 | 4.89 | 7.86 |
| PacNet (Su et al., 2019) | 1.91 | 3.20 | 5.60 | 2.48 | 4.37 | 6.60 | 2.82 | 5.01 | 8.64 | 2.40 | 4.19 | 6.95 |
| DJFR (Li et al., 2019) | 1.98 | 3.61 | 6.07 | 2.21 | **3.75** | 7.53 | 3.38 | 5.86 | 10.11 | 2.52 | 4.41 | 7.90 |
| DSRNet (Guo et al., 2018) | 2.08 | 3.26 | 5.78 | 2.57 | 4.46 | 6.45 | 3.49 | 5.70 | 9.76 | 2.71 | 4.47 | 7.30 |
| FDKN (Kim et al., 2021) | 2.21 | 3.64 | 6.15 | 2.64 | 4.55 | 7.20 | 2.63 | 4.99 | 8.67 | 2.49 | 4.39 | 7.34 |
| DKN (Kim et al., 2021) | 1.93 | 3.17 | 5.49 | 2.35 | 4.16 | 6.33 | 2.46 | 4.76 | 8.50 | 2.25 | 4.03 | 6.77 |
| Ours | **1.82** | **2.92** | **5.14** | **2.13** | 3.88 | **6.30** | **2.37** | **4.61** | **8.32** | **2.10** | **3.80** | **6.58** |

The best performance is shown in bold and second best performance is underlined

**Table 9** Comparisons on parameter numbers and model performance of depth image SR at ratio ×4 over NYU-v2 dataset

| Methods | DMSG | DJFR | DSRNet | PacNet | DKN | PMBAN | Ours |
|---|---|---|---|---|---|---|---|
| #Params | 0.33 | **0.08** | 45.49 | 0.18 | 1.16 | 25.06 | 0.13 |
| RMSE | 3.02 | 2.38 | 3.00 | 1.89 | 1.62 | 1.73 | **1.51** |

The minimum number of parameter and the best performance are shown in bold

trained on the PC with a single NVIDIA GeForce GTX 3060Ti GPU. In the training phase, these networks are optimized by the SGD optimizer over 100 epochs with a mini-batch size of 2. Furthermore, all the hidden and cell states of ConvLSTM are initialized as zero and the input $\mathbf{H}^{(0)}$ of our unfolding network is obtained by applying Bibubic upsampling over LR T1 image $\mathbf{L}$.

#### 4.3.2 Compared with SOTA Methods

The average quantitative results in ×4 and ×2 MR Image SR cases are reported in Table 10 and Table 11, respectively. As can be seen, our proposed method outperforms all the other competing methods in terms of all the indexes. Specifically, compared with the second best MINet (Feng et al., 2021), our method obtains an increase of 0.1 dB for PSNR, and 0.01 for SSIM in the ×4 ratio case. Additionally, we also make visual comparisons on some representative samples in Figs. 15 and 16, where the first row illustrates the MR Image SR results and the second row visualizes the MSE residual

**Table 10** Average performance comparison over ×4 MR Image SR

| Metrics | Bibubic | PCNN | PGN | MINet | **Ours** |
|---|---|---|---|---|---|
| PSNR↑ | 21.2330 | 32.9334 | 33.5145 | 35.1998 | **35.2928** |
| SSIM↑ | 0.6773 | 0.8933 | 0.9011 | 0.9190 | **0.9221** |

The best performance are shown in bold

**Table 11** Average performance comparison over ×2 MR Image SR

| Metrics | Bibubic | PCNN | PGN | MINet | **Ours** |
|---|---|---|---|---|---|
| PSNR ↑ | 27.3613 | 33.8824 | 35.74215 | 38.2553 | **38.3110** |
| SSIM ↑ | 0.7576 | 0.9102 | 0.9236 | 0.9449 | **0.9476** |

The best performance are shown in bold

between the generated HR MR images and the ground truth HR MR images. From Figs. 15 and 16, we can also observe that the generated HR MR images of our method have the best visual effect, and the corresponding MSE residual maps contain fewest structure information, which further supports the visual effect superiority of our method.

### 4.4 Ablation Study

To investigate the contribution of the devised modules in our proposed network, we have conducted comprehensive ablation studies on the Gaofen2 satellite dataset of the Pan-sharpening task. To be specific, the non-local cross-modalities module and persistent memory module are the two

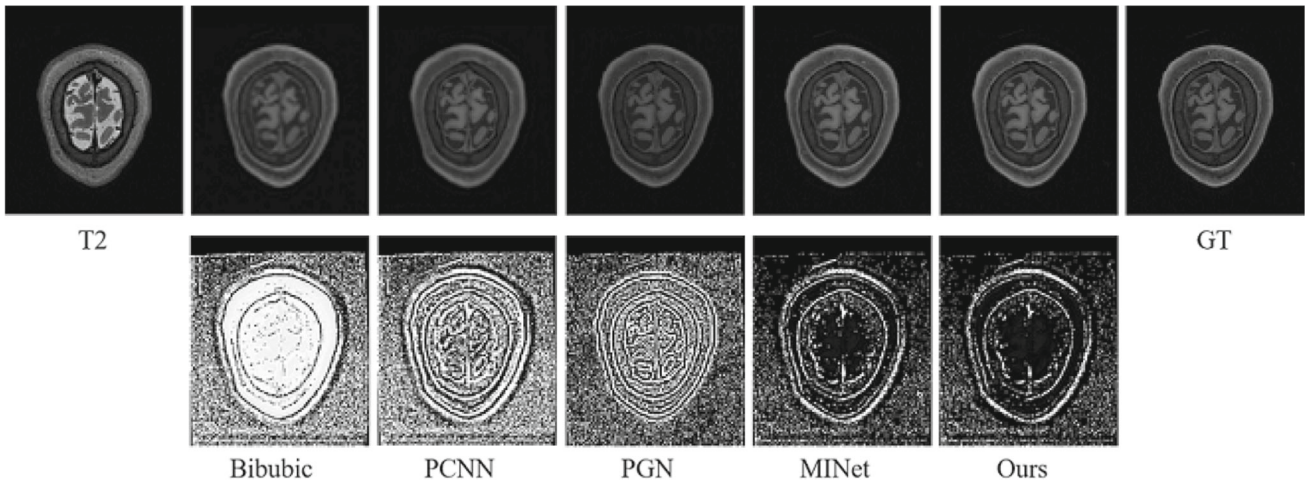



**Fig. 15** Visual comparison of $\times 2$ MRI super-resolution and T2 denotes the guidance image. Images in the last row visualizes the MSE between the generated results and the ground truth

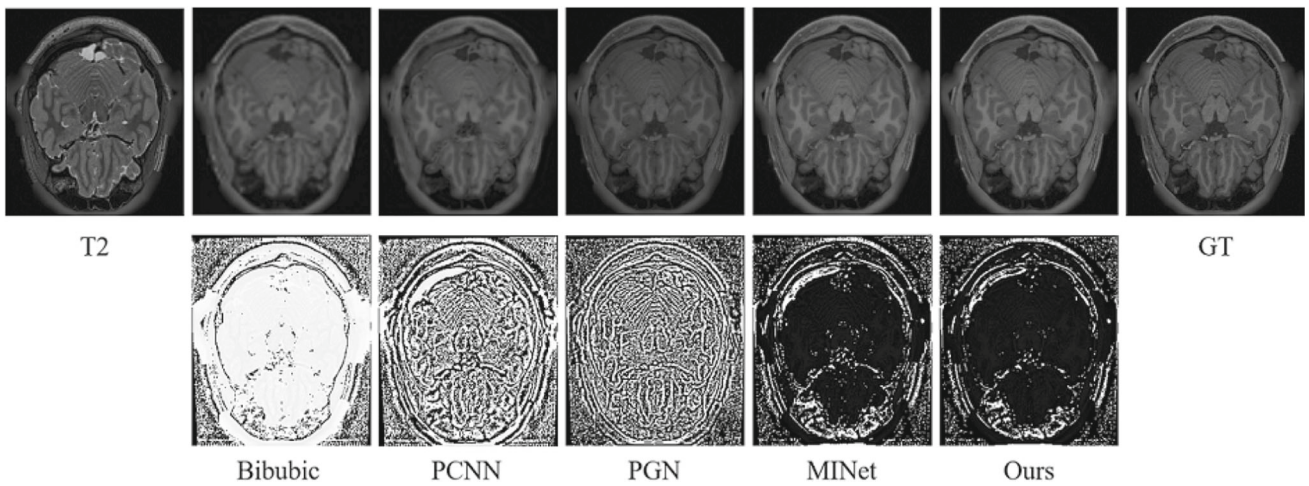



**Fig. 16** Visual comparison of $\times 4$ MRI super-resolution and T2 denotes the guidance image. Images in the last row visualizes the MSE between the generated results and the ground truth

**Table 12** Average performance comparison on the GaoFen2 datasets as the stage number increases

| Stage Number (K) | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | SCC↑ | Q ↑ | $D_\lambda$ ↓ | $D_S$ ↓ | QNR ↑ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 41.2772 | 0.9653 | 0.0249 | 1.0114 | 0.9664 | 0.7556 | 0.0616 | 0.1145 | 0.8319 |
| 2 | 41.4274 | 0.9673 | 0.0242 | 0.9834 | 0.9696 | 0.7650 | 0.0595 | 0.1106 | 0.8375 |
| 3 | 41.8058 | 0.9697 | 0.0224 | 0.9306 | 0.9737 | 0.7698 | 0.0622 | 0.1128 | 0.8329 |
| 4 | **41.8577** | **0.9697** | **0.0229** | **0.9420** | **0.9745** | **0.7740** | **0.0629** | **0.1154** | **0.8299** |
| 5 | 41.7545 | 0.9690 | 0.0226 | 0.9431 | 0.9729 | 0.7699 | 0.0600 | 0.1166 | 0.8315 |
| 6 | 41.4274 | 0.9673 | 0.0242 | 0.9834 | 0.9696 | 0.7650 | 0.0595 | 0.1106 | 0.8375 |

The best performance are shown in bold

core designs. In addition, our proposed method is developed in the iterative unfolding manner. Therefore, the studies with respect to the number of stages and the parameter sharing mechanism across stages are also conducted. Furthermore, deepening into the memory module, we also explore the effect of different-location information transmission. All the experimental results are measured by the widely-used IQA metrics, i.e., ERGAS (Alparone et al., 2007), PSNR, SSIM, SCC, Q index, SAM (J.R.H. Yuhas & Boardman, 1992), $D_\lambda$, $D_S$, and QNR.

*Effect of the stage number* The effect of stage number is first analyzed. The stage number T is set as {1, 2, 3, 4, 5, 6}.

**Table 13** Experimental results on evaluating the effect of persistent memory module and non-local cross-modalities module

| Configurations | Memory | Non-local | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | SCC↑ | Q ↑ | $D_\lambda$ ↓ | $D_S$ ↓ | QNR ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (I) | × | ✓ | 41.6287 | 0.9683 | 0.0237 | 0.9653 | 0.9727 | 0.7673 | 0.0641 | 0.1154 | 0.8287 |
| (II) | ✓ | × | 41.7665 | 0.9697 | 0.0233 | 0.9437 | 0.9742 | 0.7731 | 0.0636 | 0.1168 | 0.8279 |
| Ours | ✓ | ✓ | **41.8577** | **0.9697** | **0.0229** | **0.9420** | **0.9745** | **0.7740** | **0.0629** | **0.1154** | **0.8299** |

The best performance are shown in bold

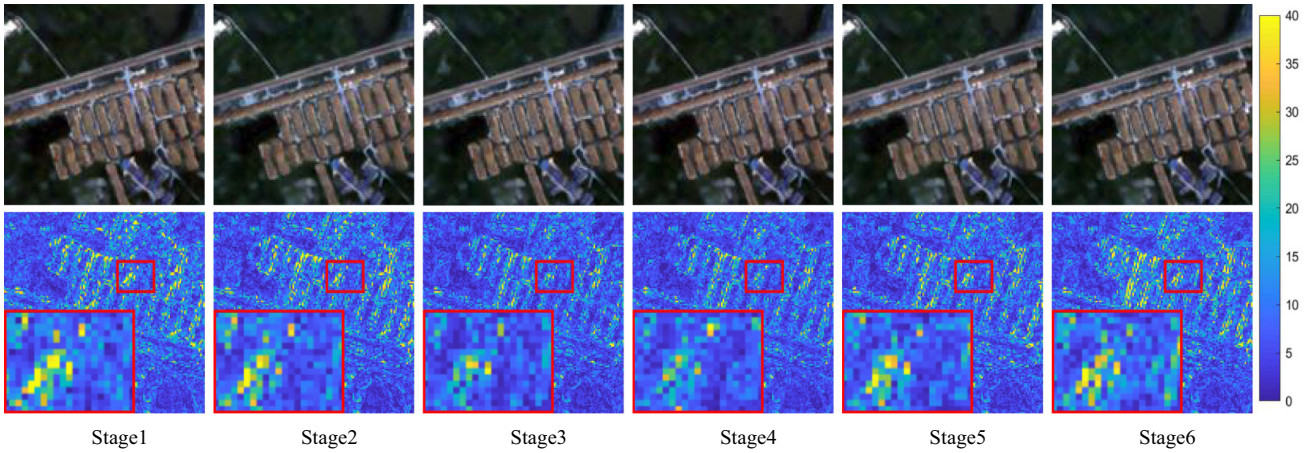



**Fig. 17** Visual comparison of different stages

The average quantitative results in terms of 8 metrics are reported in Table 12. Taking $K = 1$ as the baseline, we can see that the network performance has a continuous improvement when the stage number $K$ increases to 4. This is because our proposed network has a more powerful feature extraction ability as the stage number increases. However, when the stage number is over 4, the quantitative results show a slight decreasing trend, which may be caused by the overfitting problem due to the increasing stage. Therefore, we set $K = 4$ as the default stage number in all the experiments to balance the performance and computational complexity. Additionally, to further undertake the effect of stage number $K$, we present the representative sample generated by different model variants with stage number from $K = 1$ to $K = 6$ in Fig. 17. From Fig. 17, we can also observe that the model with stage number $K = 4$ obtains the best visual effect, which is consistent with the quantitative results of Table 12.

*Effect of parameter sharing* In this experiment, the effect of the parameters sharing mechanism across stages is verified. We take the proposed network with 4 stages as the baseline to conduct the comparison. As well recognized, the proposed network has the same network architecture in each stage. The corresponding quantitative results are reported in Table 14, from which we can observe that disabling parameter sharing is capable of improving the model performance to some extent. Besides, we can also observe that the model with parameter-sharing takes less training and inference time compared with the model without parameter-sharing. This may attribute to the increase of the model complexity when the parameter sharing mechanism is disabled. However, to achieve a good trade-off between the model parameter and model performance, we finally choose the model with parameter sharing as the default setting in all the experiments.

*Effect of persistent memory module* To explore the effectiveness of the memory module, we choose the model with memory module as the baseline and then obtain a variant of this model by eliminating the memory module from the baseline model. The quantitative results are reported in Table 13, from which it can be seen that by eliminating the memory module from the baseline, the performance in terms of all the criteria decreases. This is because memory module can reduce the information loss from the feature channel transformation and facilitate the information interaction across stages. Therefore, we can conclude that the memory module is really helpful to improve the performance.

*Effect of non-local cross-modalities module* In the section, to verify the effectiveness of the non-local cross-modalities module, we choose the model with non-local cross-modalities module as the baseline and then obtain a variant of this model by eliminating this module from the baseline model. The quantitative results are reported in Table 13. From the third and fourth row of Table 13, we can observe that by deleting the non-local cross-modalities module from the baseline, the performance with respect to all the metrics

**Table 14** Performance comparison of parameter sharing mechanism on the GaoFen2 datasets

| Parameter-sharing | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ | Training/Inference time(s) |
|---|---|---|---|---|---|
| With | 41.8577 | 0.9697 | 0.0229 | 0.9420 | 60000/0.008130 |
| Without | **42.1512** | **0.9724** | **0.0214** | **0.9042** | 61836/0.008135 |

The best performance are shown in bold

**Table 15** Performance comparison of different-location information on the GaoFen2 datasets

| Location | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ |
|---|---|---|---|---|
| Single | 41.7199 | 0.9688 | 0.0235 | 0.9461 |
| Multiple | **41.8577** | **0.9697** | **0.0229** | **0.9420** |

The best performance are shown in bold

decreases. The reason is that the non-local cross-modality module is devised to enhance the spatial resolution of the LR target image by transferring the semantically-related and structure-consistent content from the HR guidance image into the LR target image. Therefore, we can conclude that the non-local cross-modalities module is beneficial to our proposed model.

*Effect of memorizing the different-location information* As shown in Fig. 4, we add the different-layer information into the memory module at three locations. Taking the UNet at $k$-iteration as an example, $\mathbf{HP}_1^{(k)}$, $\mathbf{HP}_2^{(k)}$ and $\mathbf{U}^{(k)}$ are adopted in the output and feature spaces. In this experiment, we regard this model as baseline (denoted as Multiple), and then obtain a variant of this model (denoted as Single) by only transmitting $\mathbf{U}^{(k)}$ into the memory module. The corresponding quantitative comparison results are reported in Table 15. As can be seen, adding the memorized information at different locations will improve the performance of single location model to some extent. This is because more feature information is memorized and transformed into next iteration, and thus the information loss can be obviously alleviated. Therefore, we adopt the design of memorizing different-location information in our network. Similarly, VNet and HNet have similar implementations of memorizing the different-location information.

### 4.5 The effect of the proposed two priors

In the proposed algorithm, since the two proximal operators $\text{prox}_{\Omega_l}(\cdot)$ and $\text{prox}_{\Omega_{NL}}(\cdot)$ learnt by the deep CNNs correspond to local implicit prior and global implicit prior, respectively, we thus visualize the features from UNet that is used to learn $\text{prox}_{\Omega_l}(\cdot)$ and the features from VNet that is used to learn $\text{prox}_{\Omega_{NL}}(\cdot)$ to evaluate the effectiveness of the two priors. The learnt features are shown in Fig. 18, from which we can obviously see that the output features of the two networks are significantly different with stage increases.

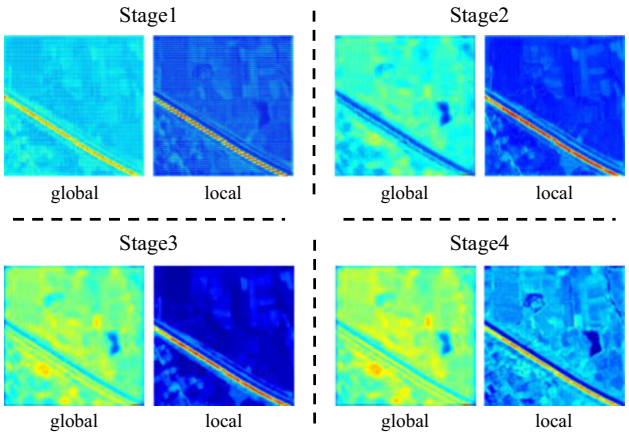



**Fig. 18** The features from UNet and VNet which correspond to local implicit prior and global implicit prior

Specifically, the UNet focuses more on local texture details while the VNet tends to capture global correlation, which further verifies the effect of the proposed local implicit prior and global implicit prior.

## 5 Limitations

As a result of the fact that our proposed method is trained on aligned data, the learned model may be less effective when applied to misaligned cases. The example shown in Fig. 19 demonstrates a severe misalignment between the HR PAN image and the LR MS image over the pan-sharpening task. Since the training and testing data have an obvious gap, it is abundantly evident that our trained network would not work well in this scenario. Consequently, our proposed model is unable to reconstruct the latent HR MS image as effective as the aligned cases, and it produces severe artifacts instead, especially in the region of aeroplane. In order to lessen the influence of the misalignment, one potential solution would be to use image matching techniques in image space or deformable convolution in feature space. This will be the focus of our future research.

## 6 Conclusion

In this work, we address GISR problem via a memory-augmented deep unfolding network which integrates both


**Fig. 19** Failure cases. The fused HR-MS image by our method is not well when the input images (PAN and LR-MS) suffer from the misalignment issue

the advantages of model-driven prior information and deep learning-based mapping capability. To facilitate the signal flow across unfolding stages, the persistent memory mechanism is introduced to augment the information representation by exploiting the Long short-term memory unit (LSTM) in the image and feature spaces. In this way, both the interpretability and information representation ability of the deep network are improved. Extensive experiments validate the superiority of our method on three representative GISR tasks, including Pan-sharpening, Depth Image SR, and MR Image SR.

Also, as verified by the experiments, the proposed maximal a posterior (MAP) estimation model is a universal method on three typical guided image super-resolution tasks, i.e., pan-sharpening, depth image super-resolution, and MRI super-resolution. Since the setting of our model is to fuse images of the same scene from two different modes, this model can also be applied to some other guided image super-resolution tasks with the same setting, i.e., hyperspectral super-resolution, digital photography image fusion, and infrared and visible image fusion. This will be one of our future work.

Additionally, our proposed method contains two fundamental designs, i.e., non-local cross-modalities learning mechanism and information-facilitating persistent memory mechanism, which can be a universal tool for other tasks. Specifically, the former mechanism (i.e., non-local cross-modalities learning mechanism) has a powerful potential of modeling the non-local self-similarity and long-range dependency over multi-modalities tasks such as reference-based image super-resolution, image fusion task and stereo vision. This mechanism can be easily embedded into the existing network architecture, and further enhance the global modeling capability. The latter mechanism (i.e., information-facilitating persistent memory mechanism) is a universal technique to alleviate the information loss issue of the deep unfolding algorithm by facilitating the information representation, and can also be easily inserted into existing deep unfolding methods.

**Acknowledgements** This work was supported by National Key Research and Development Project of China (2021ZD0110700), National Natural Science Foundation of China (62272375, 61906151, 62050194, 62037001), Innovative Research Group of the National Natural Science Foundation of China(61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology, and Project of XJTU Undergraduate Teaching Reform (20JX04Y).

## References

Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., & Bruce, L. M. (2007). Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data fusion contest. *IEEE Transactions on Geoscience and Remote Sensing, 45*(10), 3012–3021.

Bahrampour, S., Nasrabadi, N. M., Ray, A., & Jenkins, W. K. (2015). Multimodal task-driven dictionary learning for image classification. *IEEE Transactions on Image Processing, 25*(1), 24–38.

Bruna, J., Sprechmann, P., & LeCun, Y. (2015). Super-resolution with deep convolutional sufficient statistics. arXiv preprint arXiv:1511.05666

Cai, J., & Huang, B. (2020). Super-resolution-guided progressive Pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing, 59*(6), 5206–20.

Cao, X., Fu, X., Hong, D., Xu, Z., & Meng, D. (2021). Pancsc-net: A model-driven deep unfolding method for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*. https://doi.org/10.1109/TGRS.2021.3115501

Dai, S., Han, M., Xu, W., Wu, Y., & Gong, Y. (2007). Soft edge smoothness prior for alpha channel super resolution. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8

Deng, X., & Dragotti, P. L. (2019). Deep coupled ISTA network for multi-modal image super-resolution. *IEEE Transactions on Image Processing, 29*, 1683–1698.

Deng, X., & Dragotti, P. L. (2020). Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(10), 3333–48.

Diebel, J.,& Thrun, S. (2005). An application of markov random fields to range sensing. In: NIPS

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 295–307.

Dong, C., Loy, C.C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In: European conference on computer vision, Springer, pp 391–407

Dong, W., Zhang, L., Shi, G., & Li, X. (2012). Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing, 22*(4), 1620–1630.

Dong, W., Zhang, L., Lukac, R., & Shi, G. (2013). Sparse representation based image interpolation with nonlocal autoregressive modeling. *IEEE Transactions on Image Processing, 22*(4), 1382–1394. https://doi.org/10.1109/TIP.2012.2231086

Feng, C.M., Fu, H., Yuan, S., & Xu, Y. (2021). Multi-contrast mri super-resolution via a multi-stage integration network. arXiv preprint arXiv:2105.08949

Ferstl, D., Reinbacher, C., Ranftl, R., Ruether, M., & Bischof, H. (2013). Image guided depth upsampling using anisotropic total generalized variation. In: 2013 IEEE International Conference on Computer Vision, pp 993–1000, https://doi.org/10.1109/ICCV.2013.127

Geman, D., & Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*(3), 367–383.

Geman, D., & Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing, 4*(7), 932–946.

Gillespie, A. R., Kahle, A. B., & Walker, R. E. (1987). Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques - sciencedirect. *Remote Sensing of Environment, 22*(3), 343–365.

Gregor, K., & LeCun, Y. (2010). Learning fast approximations of sparse coding. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp 399–406

Guo, C., Li, C., Guo, J., Cong, R., Fu, H., & Han, P. (2018). Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing, 28*(5), 2545–2557.

Ham, B., Cho, M., & Ponce, J. (2015). Robust image filtering using joint static and dynamic guidance. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4823–4831, https://doi.org/10.1109/CVPR.2015.7299115

Ham, B., Cho, M., & Ponce, J. (2017). Robust guided image filtering using nonconvex potentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(1), 192–207.

Haydn, R., Dalke, G. W., Henkel, J., & Bare, J. E. (1982). Application of the IHS color transform to the processing of multisensor data and image enhancement. *National Academy of Sciences of the United States of America, 79*(13), 571–577.

He, K., Sun, J., & Tang, X. (2012). Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(6), 1397–1409.

He, K., Sun, J., & Tang, X. (2013). Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(6), 1397–1409.

He, R., Zheng, W. S., Tan, T., & Sun, Z. (2014). Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(2), 261–275. https://doi.org/10.1109/TPAMI.2013.102

Hirschmuller, H., & Scharstein, D. (2007). Evaluation of cost functions for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8

Hui, T.W., Loy, C.C., & Tang, X. (2016). Depth map super-resolution by deep multi-scale guidance. In: European Conference on Computer Vision, Springer, pp 353–369

J.R.H. Yuhas, A.F.G., & Boardman, J.M. (1992). Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. Proc Summaries Annu JPL Airborne Geosci Workshop pp 147–149

Jia, K., Wang, X., & Tang, X. (2012). Image transformation based on learning dictionaries across image spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(2), 367–380.

Jing, X.Y., Zhu, X., Wu, F., You, X., Liu, Q., Yue, D., Hu, R., & Xu, B. (2015). Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 695–704

Kim, B., Ponce, J., & Ham, B. (2021). Deformable kernel networks for joint image filtering. *International Journal of Computer Vision, 129*(2), 579–600.

Kim, J., Lee, J.K., & Lee, K.M. (2016a). Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1646–1654

Kim, J., Lee, J.K., & Lee, K.M. (2016b). Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1637–1645

Kingma, D.P., & Ba, J. (2017). Adam: A method for stochastic optimization. arXiv:1412.6980

Kopf, J., Cohen, M., Lischinski, D., & Uyttendaele, M. (2007a). Joint bilateral upsampling. In: ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007), vol 26

Kopf, J., Cohen, M. F., Lischinski, D., & Uyttendaele, M. (2007). Joint bilateral upsampling. *ACM Transactions on Graphics (ToG), 26*(3), 96.

Krishnan, D., & Fergus, R. (2009). Fast image deconvolution using hyper-laplacian priors. *Advances in Neural Information Processing Systems, 22*, 1033–1041.

Laben, C.A., & Brower, B.V. (2000). Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. US Patent 6,011,875

Li, Y., Huang, J., Ahuja, N., & Yang, M. (2016a). Deep joint image filtering. In: Computer Vision - 14th European Conference, ECCV 2016, Proceedings, Germany, pp 154–169, https://doi.org/10.1007/978-3-319-46493-0_10

Li, Y., Huang, J.B., Ahuja, N., & Yang, M.H. (2016b). Deep joint image filtering. In: European Conference on Computer Vision, Springer, pp 154–169

Li, Y., Huang, J. B., Ahuja, N., & Yang, M. H. (2019). Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1909–1923. https://doi.org/10.1109/TPAMI.2018.2890623

Liao, W., Xin, H., Coillie, F.V., Thoonen, G., & Philips, W. (2017). Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter. In: Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing

Liu, D., Wang, Z., Wen, B., Yang, J., Han, W., & Huang, T. S. (2016). Robust single image super-resolution via deep networks with sparse prior. *IEEE Transactions on Image Processing, 25*(7), 3194–3207.

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

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., & Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3550–3557

Lu, S., Ren, X., & Liu, F. (2014). Depth enhancement via low-rank matrix completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3390–3397

Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., & Wang, G. (2020). Multi-contrast super-resolution MRI through a progressive network. *IEEE Transactions on Medical Imaging, 39*(9), 2738–2749.

Mallat, S., & Yu, G. (2010). Super-resolution with sparse mixing estimators. *IEEE Transactions on Image Processing, 19*(11), 2889–2900.

Mao, X., Shen, C., & Yang, Y. B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in Neural Information Processing Systems, 29*, 2802–2810.

Marivani, I., Tsiligianni, E., Cornelis, B., & Deligiannis, N. (2020). Multimodal deep unfolding for guided image super-resolution. *IEEE Transactions on Image Processing, 29*, 8443–8456.

Masi, G., Cozzolino, D., Verdoliva, L., & Scarpa, G. (2016). Pansharpening by convolutional neural networks. *Remote Sensing, 8*(7), 594.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A.Y. (2011). Multimodal deep learning. In: IEEE International Conference on Machine Learning (ICML)

Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., de Marvao, A., Cook, S., O'Regan, D., & Rueckert, D. (2016). Multi-input cardiac image super-resolution using convolutional neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, pp 246–254

Park, J., Kim, H., Tai, Y.W., Brown, M.S., & Kweon, I. (2011). High quality depth map upsampling for 3d-tof cameras. In: 2011 Inter-

national Conference on Computer Vision, pp 1623–1630, https://doi.org/10.1109/ICCV.2011.6126423
Paszke, A., Gross, S., Massa, F., Lerer, A., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library
Pham, C.H., Ducournau, A., Fablet, R., & Rousseau, F. (2017). Brain mri super-resolution using deep 3d convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, pp 197–200
Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *Siam J Control Optim, 14*(5), 877–898.
Sanchez-Beato, A., & Pajares, G. (2008). Noniterative interpolation-based super-resolution minimizing aliasing in the reconstructed image. *IEEE Transactions on Image Processing, 17*(10), 1817–1826.
Scharstein, D., & Pal, C. (2007). Learning conditional random fields for stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8
Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision, 47*(1), 7–42.
Shen, X., Yan, Q., Xu, L., Ma, L., & Jia, J. (2015). Multispectral joint image restoration via optimizing a scale map. *IEEE transactions on pattern analysis and machine intelligence, 37*(12), 2518–2530.
Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgb-d images. In: Proceedings of the European Conference on Computer Vision, pp 746–760
Song, J., Chen, B., & Zhang, J. (2021). Memory-augmented deep unfolding network for compressive sensing. In: ACM MM
Song, P., Deng, X., Mota, J. F., Deligiannis, N., Dragotti, P. L., & Rodrigues, M. R. (2019). Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries. *IEEE Transactions on Computational Imaging, 6*, 57–72.
Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., & Kautz, J. (2019). Pixel-adaptive convolutional neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 11158–11167, https://doi.org/10.1109/CVPR.2019.01142
Sun, B., Ye, X., Li, B., Li, H., Wang, Z., & Xu, R. (2021). Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7788–7797, https://doi.org/10.1109/CVPR46437.2021.00770
Sun, J., Xu, Z., & Shum, H.Y. (2008). Image super-resolution using gradient profile prior. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8
Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3147–3155
Timofte, R., De Smet, V., & Van Gool, L. (2013). Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1920–1927
Timofte, R., De Smet, V., & Van Gool, L. (2014). A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian Conference on Computer Vision, Springer, pp 111–126
Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In: Proceedings of the IEEE International Conference on Computer Vision, IEEE, pp 839–846
Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., Restaino, R., & Wald, L. (2014). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing, 53*(5), 2565–2586.
Wald, L., Ranchin, T., & Mangolini, M. (1997). Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing, 63*, 691–699.
Wang, J., Chen, Y., Wu, Y., Shi, J., & Gee, J. (2020). Enhanced generative adversarial network for 3d brain mri super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 3627–3636
Wang, S., Zhang, L., Liang, Y., & Pan, Q. (2012). Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2216–2223
Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7794–7803
Wu, H., Zheng, S., Zhang, J., & Huang, K. (2018a). Fast end-to-end trainable guided filter. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1838–1847
Wu, H., Zheng, S., Zhang, J., & Huang, K. (2018b). Fast end-to-end trainable guided filter. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1838–1847, https://doi.org/10.1109/CVPR.2018.00197
Xu, S., Zhang, J., Zhao, Z., Sun, K., Liu, J., & Zhang, C. (2021). Deep gradient projection networks for pan-sharpening. In: CVPR, pp 1366–1375
Yang, J., Wright, J., Huang, T., & Ma, Y. (2008). Image super-resolution as sparse representation of raw image patches. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8
Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing, 19*(11), 2861–2873.
Yang, J., Wang, Z., Lin, Z., Cohen, S., & Huang, T. (2012). Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing, 21*(8), 3467–3478.
Yang, J., Lin, Z., & Cohen, S. (2013). Fast image super-resolution based on in-place example regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1059–1066
Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., & Paisley, J. (2017). Pannet: A deep network architecture for pan-sharpening. In: IEEE International Conference on Computer Vision, pp 5449–5457
Ye, X., Sun, B., Wang, Z., Yang, J., Xu, R., Li, H., & Li, B. (2020). Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing, 29*, 7427–7442. https://doi.org/10.1109/TIP.2020.3002664
Yuan, Q., Wei, Y., Meng, X., Shen, H., & Zhang, L. (2018). A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(3), 978–989.
Zeng, K., Zheng, H., Cai, C., Yang, Y., Zhang, K., & Chen, Z. (2018). Simultaneous single-and multi-contrast super-resolution for brain mri images based on a convolutional neural network. *Computers in Biology and Medicine, 99*, 133–141.
Zhang, K., Gool, L.V., & Timofte, R. (2020). Deep unfolding network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3217–3226
Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018a). Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 286–301
Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018b). Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2472–2481

Zhou, M., Fu, X., Huang, J., Zhao, F., Liu, A., & Wang, R. (2022). Effective pan-sharpening with transformer and invertible neural network. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–15. https://doi.org/10.1109/TGRS.2021.3137967

Zhou, M., Huang, J., Fang, Y., Fu, X., & Liu, A. (2022b). Pan-sharpening with customized transformer and invertible neural network. In: Thirty-Six AAAI Conference on Artificial Intelligence

Zhou, M., Yan, K., Huang, J., Yang, Z., Fu, X., & Zhao, F. (2022c). Mutual information-driven pan-sharpening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1798–1808

Zhuang, Y.T., Wang, Y.F., Wu, F., Zhang, Y., & Lu, W.M. (2013). Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: Twenty-Seventh AAAI Conference on Artificial Intelligence