Guided Deep Decoder: Unsupervised Image Pair Fusion

Tatsumi Uezato[1[0000−0002−8264−201X]], Danfeng Hong[2,3[0000−0002−3212−9584]], Naoto Yokoya[4,1[0000−0002−7321−4590]], and Wei He[1[0000−0003−3410−0643]]

[1] RIKEN AIP, Tokyo, Japan
[2] German Aerospace Center, Wessling, Germany
[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
[4] The University of Tokyo, Tokyo, Japan
{tatsumi.uezato,naoto.yokoya,wei.he}@riken.jp, {danfeng.hong}@dlr.de

Abstract. The fusion of input and guidance images that have a tradeoff in their information (e.g., hyperspectral and RGB image fusion or pansharpening) can be interpreted as one general problem. However, previous studies applied a task-specific handcrafted prior and did not address the problems with a unified approach. To address this limitation, in this study, we propose a guided deep decoder network as a general prior. The proposed network is composed of an encoder-decoder network that exploits multi-scale features of a guidance image and a deep decoder network that generates an output image. The two networks are connected by feature refinement units to embed the multi-scale features of the guidance image into the deep decoder network. The proposed network allows the network parameters to be optimized in an unsupervised way without training data. Our results show that the proposed network can achieve state-of-the-art performance in various image fusion problems.

Keywords: Deep Image Prior, Deep Decoder, Image Fusion, Hyperspectral Image, Super-resolution, Pansharpening.

1 Introduction

Some image fusion tasks address the fusion of image pairs in the same modality. The tasks consider a pair of images that capture the same region but have a tradeoff between the two images (Fig. 1). For example, a low spatial resolution hyperspectral (LR-HS) image has greater spectral resolution at lower spatial resolution [39]. However, an RGB image acquires much lower spectral resolution at higher spatial resolution. Likewise, panchromatic and multispectral (MS) images have a tradeoff between spatial and spectral resolution [29]. No-flash images capture ambient illumination, but are very noisy, while flash images capture artificial light, but are less noisy [23]. Image fusion enables an image that overcomes the tradeoff to be generated. Hyperspectral super-resolution or pansharpening aims to generate a high resolution (HR) HS or MS image. The denoising of a

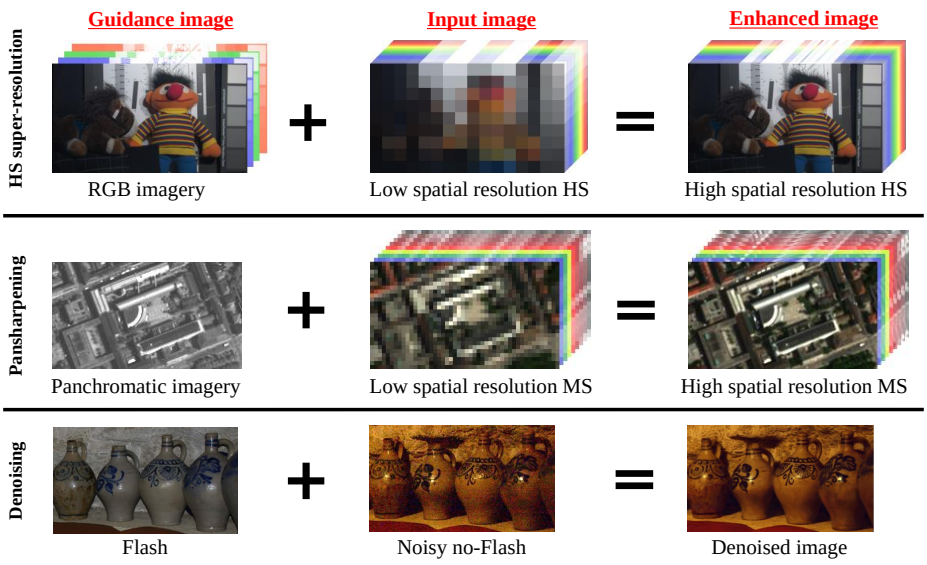

Fig. 1: Illustration of image pair fusion of the same modality.

no-flash image with a flash image can be also interpreted as a special case of image fusion.

Although these tasks share a common goal (i.e., enhancing input images with the help of guidance images), the tasks have been studied independently. This occurs because a different handcrafted prior is considered to incorporate the specific property of an output image. In HS super-resolution, a prior exploiting the low-rankness of HS has been extensively used [40, 7, 18]. In pansharpening, a prior representing a spatial smoothness has been considered [22]. The denoising task assumes that the spatial structure of a restored image is similar to that of a guidance image [23]. While these handcrafted priors share the same goal, the priors need to be designed for each task to exploit the specific properties of data. It is highly desirable to develop a prior applicable to various image fusion problems.

Deep learning (DL) approaches avoid the assumption of explicit priors for each specific task. Although network architectures themselves need to be handcrafted, properly designed network architectures have shown to solve various problems [25, 14]. Most DL approaches rely on training data. However, for pansharpening and hyperspectral super-resolution, it is difficult to collect a large size of training data including reference (i.e., HR-HS or HR-MS) because of the cost or hardware limitation. Thus, previous studies [36, 26] have frequently used synthetic data for training, which may have limited generalization performance. In addition, different sensors provide different spectral response functions. Networks trained on data acquired by a particular sensor may not work well on new data acquired by a different sensor.

A natural question arises: is it possible to use DL approaches without training data? Ulyanov *et al.* [28] have shown that network architectures have inductive bias and can be used as deep image prior (DIP) without any training data. This intriguing property of DIP has been successfully used for various problems [12, 38, 27]. In [28], the guided denoising task of flash and no-flash image pair has been addressed using a no-flash image as an input and a flash image as an output. Although this approach can be potentially used to address the problems shown in Fig. 1, the network architecture does not fully exploit the semantic features or image details of a guidance image. It is still unclear how the network architecture is conditioned on the features of a guidance image. Although DIP has great potential, the uncertainties limit DIP to achieve state-of-the-art (SOTA) performance in various image fusion problems.

As discussed above, previous studies face two major problems (task-specific handcrafted priors and requirement of training data) to address various image fusion problems in a unified framework. In this study, we propose a new network architecture, called a guided deep decoder (GDD), that overcomes the problems and can achieve SOTA performance in different image fusion problems. Specifically, the proposed network architecture is composed of two networks where one encoder-decoder network is designed to extract multi-scale semantic features from a guidance image, while another deep decoder network generates an output image from random noise. The two networks are connected by feature refinement units incorporating attention gates to embed the multi-scale features of the guidance image into the deep decoder network.

The contributions of this paper are as follows. (1) We propose a new unsupervised DL method that does not require training data and can be adapted to different image fusion tasks in a unified framework. We achieve SOTA results for various image fusion problems. (2) We propose a new network architecture as a regularizer for unsupervised image fusion problems. The attention gates used in the proposed architecture guide the generation of an output image using the multi-scale semantic features from a guidance image. The guidance of the multi-scale features can lead to an effective regularizer for an ill-posed optimization problem.

2 Related work

Most of the previous works have independently addressed one of the image fusion problems shown in Fig. 1, although the common goal is to generate an image that overcomes the tradeoff. This study focuses on the data acquired in the same modality and is different from the image fusion problems of different modalities where the sensor captures different physical quantities (e.g., fusion of RGB images and depth maps [20]). To address the ill-posed fusion problems, similar approaches have been developed for different image fusion tasks.

Classical approach: The classical approach is to specifically design a handcrafted prior for each task. For example, handcrafted priors exploiting the low-rankness or sparsity of HS have been developed for HS and MS image fusion

problems [40, 16, 17, 34]. In panchromatic and MS image fusion, the handcrafted priors, which assume that the spatial details of PAN are similar to those of MS, have been widely used [19, 22, 10, 6]. In addition, flash and no-flash image fusion uses a prior that promotes similar spatial details between the paired image [23]. The classical approach can reconstruct an enhanced image without any training data by explicitly assuming prior knowledge. However, the priors designed for a specific task may not be effective when they are applied to other tasks. In addition, an optimization method needs to be tailored for a different prior.

Supervised DL approach: DL methods that use training data have recently achieved SOTA performance in different image fusion problems. DL methods are usually built upon a popular network (e.g., [25, 14]). In the HS and RGB image fusion, DL methods use LR-HS and RGB images as an input and an HR-HS image as an output and learn the mapping function between the inputs and the output [36, 8]. Similarly, in pansharpening, the methods consider panchromatic and LR-MS images as an input and HR-MS as an output and learn the mapping function [26, 35, 37]. As long as training data are available, DL methods can be potentially applied to different image fusion problems in a unified framework by slightly changing the network architecture or the loss function. However, it may be difficult to acquire training data, including reference data, for HS or MS images because of the cost or hardware limitation.

Unsupervised DL approach: To bridge the gap between the classical and supervised DL approaches, an unsupervised DL approach has been considered in some studies. The unsupervised DL methods have been developed to address the HS and RGB image fusion problem [24, 11]. In [24, 11], the network architecture has been specifically designed to exploit the property of the HS image and different handcrafted priors have been combined to achieve optimal performance. However, it may not achieve SOTA performance in other tasks because of the specifically designed network and handcrafted priors. DIP that can apply DL in an unsupervised way has been recently developed by [28] and has been applied for a variety of problems [12, 38, 27]. Although DIP can be potentially applied for various image fusion problems, it has not been explored yet. The simple application of DIP cannot achieve SOTA performance in different image fusion tasks, which is shown in the following experiments. Our study borrows the idea of DIP and proposes a robust network architecture that achieves SOTA performance in these tasks.

3 Methodology

3.1 Problem formulation

Let us denote a low resolution or noisy input image $\mathbf{Y} \in \mathbb{R}^{C \times w \times h}$ and a guidance image $\mathbf{G} \in \mathbb{R}^{c \times W \times H}$ where C, W, and H represent the number of channels, the image width, and the image height, respectively. When considering HS super-resolution or pansharpening, $w \ll W$, $h \ll H$, and $c \ll C$. In the unsupervised image fusion problem, the corresponding output $\mathbf{X} \in \mathbb{R}^{C \times W \times H}$ can be estimated

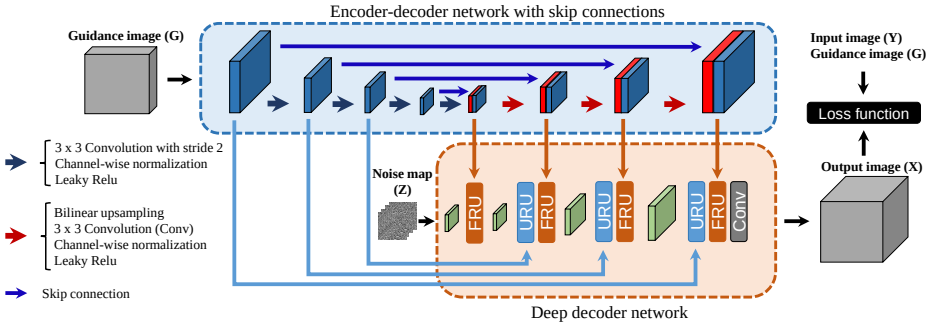

Fig. 2: The structure of a guided deep decoder. The semantic features are extracted from the guidance image by the U-net like encoder-decoder network. The blue layers represent the features of the encoder. The red layers represent the features of the decoder. The green layers represent the features of the deep decoder network. The semantic features of G are used to guide the features of the deep decoder in the upsampling and feature refinement units (URU and FRU).

by solving the following optimization problem:

$$\min_{\mathbf{X}} \mathcal{L}\left(\mathbf{X}, \mathbf{Y}, \mathbf{G}\right) + \mathcal{R}\left(\mathbf{X}\right), \tag{1}$$

where $\mathcal{L}$ is a loss function that is different for each task. Because the problem is ill-posed, existing methods commonly add the handcrafted regularization term $\mathcal{R}$. However, the task-specific regularization term (e.g., low-rank property of HS images) cannot be easily applied to other tasks. Instead of using the handcrafted regularization terms, DIP estimates $\mathbf{X}$ using a convolutional neural network (CNN)-based mapping function as:

$$\mathbf{X} = f_\theta\left(\mathbf{Z}\right), \tag{2}$$

where f_θ represents the mapping function with the network parameters θ, $\mathbf{Z}$ is the input representing the random code tensor. The optimization problem can be rewritten as:

$$\min_{\theta} \mathcal{L}\left(f_\theta\left(\mathbf{Z}\right), \mathbf{Y}, \mathbf{G}\right). \tag{3}$$

In this formulation, only one input image $\mathbf{Y}$ and a guidance image $\mathbf{G}$ are used for the optimization problem; thus, training data are *not* required. $\mathbf{X}$ is regularized by the implicit prior of the network architecture. Different types of architectures can lead to different regularizers. The architecture that effectively incorporates multi-scale spatial details and semantic features of the guidance image can be a powerful regularizer for the optimization problem. In the following section, we propose a new architecture, called the guided deep decoder, as a regularizer that can be used for various image fusion problems.

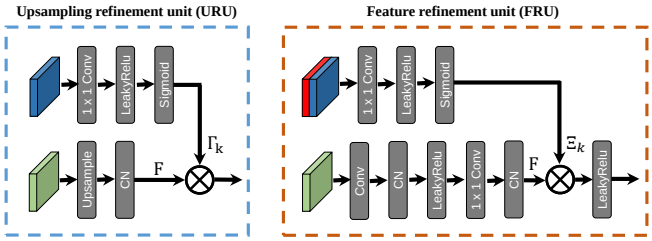

Fig. 3: The structure of upsampling and feature refinement units.

3.2 Guided deep decoder (GDD)

GDD is composed of an encoder-decoder network with skip connections and a deep decoder network, as shown in Fig. 2. The encoder-decoder network is similar to the architecture of U-net [25] and produces the features of a guidance image at multiple scales. The multi-scale features represent hierarchical semantic features of the guidance image from low to high levels. The semantic features are used to guide the parameter estimation in the deep decoder. Let $\mathbf{\Gamma}_k$ denote the features of the encoder at the kth scale, $\mathbf{\Xi}_k$ denotes the kth-scale features in the decoder part of the encoder-decoder network. The mapping function is conditioned on the multi-scale features as $f_\theta\left(\mathbf{Z}|\,\mathbf{\Gamma}_1,\cdots,\mathbf{\Gamma}_K,\mathbf{\Xi}_1,\cdots,\mathbf{\Xi}_K\right)$. The multi-scale features are incorporated in the deep decoder by the two proposed units shown in Fig. 3.

Upsampling refinement unit (URU). Upsampling is a vital part of DIP [4]. Bilinear or nearest neighbor upsampling promotes piecewise constant patches or smoothness across all channels [15]. However, the prior is too strong to recover exact spatial structures or boundaries of an image. Although this problem is alleviated using skip connections, the spatial details of a guidance image are still lost in the features of the decoder. URU incorporates an attention gate for weighting the features derived after upsampling and channel-wise normalization (CN) in the deep decoder. The features from the guidance image are gated by a 1×1 convolution (Conv), a leaky rectified linear unit (LeakyRelu), and a sigmoid activation layer (Sigmoid) to preserve the spatial locality of the features and generate the conditional weights. Given the features of the deep decoder $\mathbf{F}$, the transformation is carried out as:

$$\mathbf{URU}\left(\mathbf{F}|\mathbf{\Gamma}_k\right) = \mathbf{F} \otimes \mathbf{\Gamma}_k, \tag{4}$$

where $\otimes$ represents the element-wise multiplication. Note that the dimensions of $\mathbf{F}$ and $\mathbf{\Gamma}_k$ are the same at each scale. Both channel-wise and spatial-wise conditional weights are considered in URU.

Feature refinement unit (FRU). FRU is different from URU in that the features of the deep decoder are weighted by the high-level semantic features

of the guidance image. FRU promotes the semantic alignment with the features of the guidance image, while URU promotes similar spatial locality. Using an attention gate, the high-level features are gated by a 1×1 convolution, a leaky rectified linear unit, and a sigmoid activation layer to generate the conditional weights. FRU transforms the features of the deep decoder as follows:

$$\mathbf{FRU}\left(\mathbf{F}|\mathbf{\Xi}_k\right) = \mathbf{F} \otimes \mathbf{\Xi}_k. \tag{5}$$

Note that the dimensions of $\mathbf{F}$ and $\mathbf{\Xi}_k$ are the same at each scale. The features of the deep decoder are weighted in URU and FRU, which leads to a deep prior that can more explicitly exploit the spatial details or semantic features of the guidance image than DIP.

3.3 Loss Function

The loss function is different for each task. In this section, the loss functions used for HS super-resolution, pansharpening, and denoising are discussed.

HS super-resolution. When fusing RGB and HS images, the loss function is usually designed to preserve the spectral information from the HS image while keeping the spatial information from the RGB image. For simplicity, the matrix forms of $\mathbf{X}, \mathbf{Y}, \mathbf{G}$ are denoted as $\tilde{\mathbf{X}} \in \mathbb{R}^{C \times WH}$, $\tilde{\mathbf{Y}} \in \mathbb{R}^{C \times wh}$, and $\tilde{\mathbf{G}} \in \mathbb{R}^{c \times WH}$, respectively. Given the estimated HR-HS $\tilde{\mathbf{X}}$, the loss function can be defined as:

$$\mathcal{L}\left(\mathbf{X}, \mathbf{Y}, \mathbf{G}\right) = \mu \|\tilde{\mathbf{X}}\mathbf{S} - \tilde{\mathbf{Y}}\|_F^2 + \|\mathbf{R}\tilde{\mathbf{X}} - \tilde{\mathbf{G}}\|_F^2, \tag{6}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{S}$ is the spatial downsampling with blurring and $\mathbf{R}$ is the spectral response function that integrates the spectra into R, G, B channels. The first term encourages the spectral similarity between the spatially downsampled $\mathbf{X}$ and $\mathbf{Y}$. The second term encourages the spatial similarity between the spectrally downsampled $\mathbf{X}$ and $\mathbf{G}$. μ is a scalar controlling the balance between the two terms. The loss function has been widely used with the handcrafted priors in the HS super-resolution [18, 40] because the optimization problem is highly ill-posed. Our approach differs from those used in previous studies because it uses GDD as a regularizer.

Pansharpening. Like HS super-resolution, pansharpening also considers two terms that balance the tradeoff between spatial and spectral information. Although the first term in (6) can be also used for the loss function of pansharpening, the second term may not be effective. This is because the spectral response function of the pansharpening image may partially cover the spectral range captured by the MS image. Thus, the second term cannot effectively measure the spatial similarity between panchromatic and MS images. To address the problem, the second term measuring the spatial similarity is defined as follows:

$$\mathcal{L}\left(\mathbf{X}, \mathbf{Y}, \mathbf{G}\right) = \mu \|\tilde{\mathbf{X}}\mathbf{S} - \tilde{\mathbf{Y}}\|_F^2 + |\mathbf{D}\nabla\tilde{\mathbf{X}} - \nabla\tilde{\mathbf{G}}|, \tag{7}$$

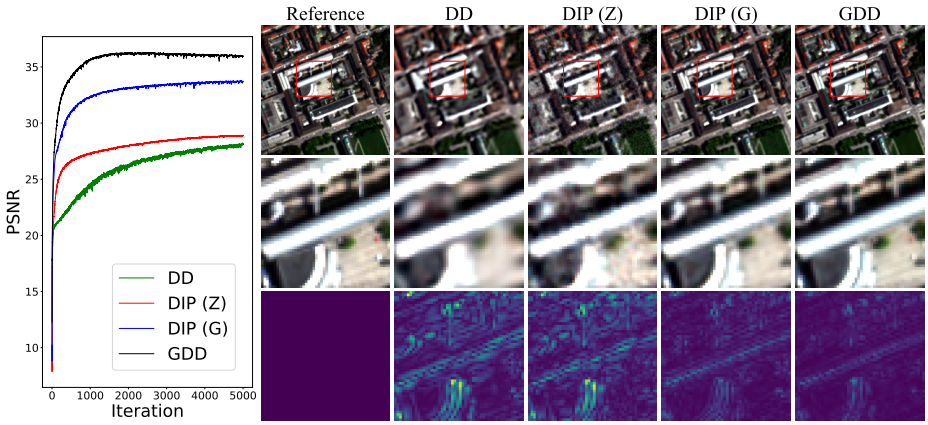

Fig. 4: Comparison of DD, DIP, and GDD. The left figure shows PSNR at different iterations. The right figure shows the images derived at the 5000 iterations. From top to bottom, RGB images, enlarged RGB images, and the error maps of the compared methods.

where $\tilde{\mathbf{Y}}$ is the MS image, $\tilde{\mathbf{G}}$ is the panchromatic image expanded to the same number of bands of $\tilde{\mathbf{X}}$, $\nabla\tilde{\mathbf{X}}$ is the image gradient of $\tilde{\mathbf{X}}$, $\nabla\tilde{\mathbf{G}}$ is the image gradient of $\tilde{\mathbf{G}}$, $|\cdot|$ is the l_1 norm, and $\mathbf{D}$ is the diagonal matrix to weight each channel of $\nabla\tilde{\mathbf{X}}$ so that the magnitude of $\tilde{\mathbf{X}}$ is scaled to that of $\nabla\tilde{\mathbf{G}}$. Note that $\mathbf{D}$ can be learned with other parameters within the GDD optimization framework. The l_1 norm is chosen because this norm more explicitly encourages the edges of the output and guidance images to be similar than other norms (e.g., l_2 norm). The first term encourages the spectral similarity while the second term promotes the spatial similarity. A similar loss function has been also explored in [5].

Denoising. For the denoising of the no-flash image, the following loss function was used:

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = \|\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}\|_F^2, \tag{8}$$

where $\tilde{\mathbf{Y}}$ is the no-flash image. Only $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$ are considered in the loss function. $\tilde{\mathbf{G}}$ is considered only in the network architecture because in the detail transfer of the flash and no-flash images, the spatial structures or colors are not necessarily consistent [23]. To fairly compare the results derived by DIP [28], we adopt the same loss function.

Different handcrafted priors are usually considered with task-specific loss functions. As a result, an optimization framework can be also different for each task. Our approach is different from the previous studies in that GDD is used as a common prior for all of the tasks in a unified optimization framework.

4 Comparison between DD, DIP, and GDD

In this section, we show the comparison between a deep decoder (DD), DIP, and GDD to discuss how GDD outperforms the compared methods. Extensive experiments, including other applications, are shown in the following section. Fig. 4 shows peak signal-to-noise ratio (PSNR) at different iterations. DD uses a tensor representing random noise as an input. DD corresponds to the deep decoder part in GDD. DD is considered for comparison to validate whether the features guided by the encoder-decoder network are really useful. DIP (Z) represents the deep image prior that uses a random tensor as an input, while DIP (G) uses a guidance image (i.e., panchromatic imagery) as an input in the encoder-decoder network. Because DD considers only the decoder part, the information lost in the process of upsampling cannot be recovered. DIP(Z) can use the features derived by a skip connection as a bias term and try to compensate for the lost information. This led to slightly better results of DIP(Z). GDD and DIP (G) that incorporate the guidance image produced high PSNR at early iterations. This shows that the use of the guidance image leads to the high quality of the HR-MS image at fewer iterations. Although both GDD and DIP (G) use the guidance image, GDD considerably outperformed DIP in terms of PSNR. Fig. 4 also shows the RGB images of the reconstructed images, the enlarged RGB, and the corresponding error maps. The enlarged RGB image derived from DD is blurred. The image derived by DIP (Z) is also blurred and the texture is not correctly recovered. In the highly ill-posed optimization problem, the deep prior that does not incorporate the guidance image cannot produce satisfactory results. DIP (G) performs better than DD or DIP (Z). However, the small objects or boundaries of the image are missing in the reconstructed image. GDD preserved the small objects or boundaries more explicitly than DIP (G), which led to smaller errors. In addition, GDD produced smaller errors in the homogeneous regions of the objects.

Reasons why GDD is a good regularizer. We argue that GDD works as a better regularizer than DIP (G) for the following two reasons:
1. **Upsampling refinement**: The bilinear upsampling used in DIP and GDD causes a strong bias to promote piecewise smoothness and tends to wash away the small objects or boundaries. GDD differs from DIP because it uses an attention gate to weight the features derived by the upsampling. The attention gate enables the small objects or boundaries to be aware by the conditional weights shown in Fig. 5. Owing to attention gates, GDD can reconstruct spatial details.
2. **Feature refinement at multiple scales**: DIP uses the guidance image as an input in the hourglass architecture. In DIP, the features of each layer in the decoder part of the architecture are conditioned using only the features of the previous layer. GDD enables the features of each layer in the decoder part to be conditioned on the semantic features from the guidance image at multiple scales. The attention gates at multiple scales emphasize salient features within each layer, leading to the semantic alignment between the output image and the guidance image.

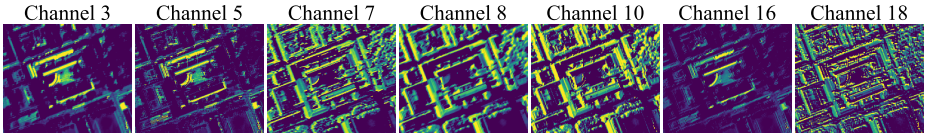

Fig. 5: Examples of the conditional weights of different channels used in the attention gates.

5 Experiments

In this section, we show how GDD works as a regularizer for different image fusion problems. Because of the limited space, only the selected results are shown in the main document. Additional results are shown in the supplementary material. The network architecture of GDD has been fixed for all of the following experiments to validate the robustness of GDD as a regularizer. It is possible to carefully tune the network architecture for each task. However, we believe that the fixed network architecture that works well for different tasks is more important than a carefully tuned architecture that obtains the best performance only for a specific task. In the following experiments, DIP used the guidance image as an input and the same loss function with GDD for fair comparison.

5.1 Hyperspectral super-resolution

Dataset. The CAVE dataset[5] was chosen for the experiments because it has been extensively used to evaluate HS super-resolution methods [7, 24, 11, 36]. The CAVE dataset consists of HR-HS images that were acquired in 32 indoor scenes with controlled illumination. Each HR-HS image has the spatial size of 256×256 with 31 bands representing the reflectance of materials in the scene. We followed the experimental setup of [36], *i.e.*, the generation of the LR-HS image from the HR-HS image by averaging over 32×32 pixel blocks and the generation of the RGB image by spectral downsampling on the basis of the spectral response function. The proposed GDD does not require training data. However, for fair comparison with the supervised DL method, we chose 12 images for the test, and the rest of the images were used for training as done in [36].

Compared methods. The compared SOTA methods include the matrix/tensor related methods (CNMF [40], BSR [1], NSSR [9] and NLSTF [7]), the supervised DL method (MHF [36]), and the unsupervised DL methods (UDL [11], uSDN [24] and DIP [28]). Among all methods, only MHF required training data.

To quantitatively validate the results, four different criteria were used. The criteria are the root mean square error (RMSE), spectral angle (SA), structural similarity (SSIM [33]), and the relative dimensionless global error in synthesis (ERGAS [31]).

[5] `http://www1.cs.columbia.edu/CAVE/databases/`

Fig. 6: First row: reference and RGB images of the reconstructed HS. The selected results are from *chart and staffed toy* in the CAVE data. Second row: The corresponding error maps.

Results. Table 1 shows the average results of all test images. The performance of BSR and uSDN was worse than those of other methods because the two methods do not assume that the downsampling matrix is available *a priori*. GDD outperformed other unsupervised HS super-resolution methods and was even competitive with the trained DL method (i.e., MHF). This shows that the proposed network architecture is an effective regularizer for the HS super-resolution problem. Fig. 6 shows the RGB images of the reconstructed HS images and the error maps. In general, GDD produced lower errors than other methods. The noticeable difference between DIP and GDD is that the errors of DIP are significantly larger at the edges of the image than those of GDD. This implies that GDD properly incorporates the spatial details or semantic features of the guidance image, leading to the edge-preserving image.

Table 1: Quantitative results of different metrics on the CAVE dataset. ↓ shows lower is better while ↑ shows higher is better.

	CNMF	BSR	NLSTF	NSSR	UDL	uSDN	MHF	DIP	GDD
RMSE↓	3.4557	5.2030	2.9414	2.4247	2.7971	4.9289	2.0827	3.1589	**2.0213**
ERGAS↓	0.5347	0.7318	0.4144	0.3696	0.3650	0.7723	0.3062	0.4597	**0.3041**
SA↓	7.0801	13.1719	8.9825	7.4138	6.9816	12.4995	6.0100	7.6734	**5.5740**
SSIM↑	0.9760	0.9524	0.9805	0.9770	0.9733	0.9385	**0.9874**	0.9621	0.9869

5.2 Pansharpening

Dataset. Four different image scenes covering agriculture, urban, forest or mixtures of these were chosen for the experiments. The images were acquired by the WorldView-2. Each MS image is composed of 8 bands representing spectral reflectance. The spatial resolution of the MS image is 2 m while that of the panchromatic image is 0.5 m. Each panchromatic image has one band that partially covers the spectral range of the MS image. Synthetic MS and panchromatic images were generated by spatially downsampling the original resolution MS and

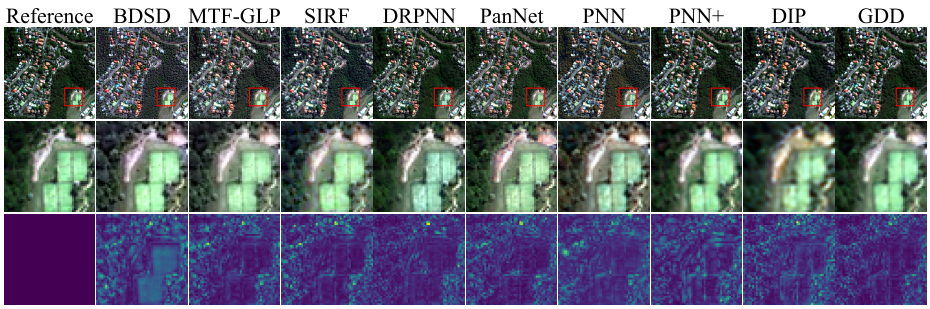

Fig. 7: First row: RGB images of the pansharpened MS images. Second row: The enlarged RGB images. Third row: The corresponding error maps.

panchromatic images by the factor of 4. Bicubic downsampling was used. The original resolution MS image was used as reference data. This is the common approach called Wald's protocol [30] to generate reference data because reference data (i.e., HR-MS image) are not available [10, 37].

Compared methods. The compared SOTA methods include three unsupervised pansharpening methods (BDSD [13], MTF-GLP [29], SIRF [5]) and four supervised DL methods (DRPNN [35], PanNet [37], PNN [21], PNN+ [26]). The supervised DL methods achieved the SOTA performance. However, the generalization performance of the supervised DL methods is still limited if training data are acquired by a different sensor or in different regions. Training data must be carefully prepared for the supervised DL methods. In this study, we divided each image scene into training and test data acquired by the same sensor. This produces a favorable condition for the supervised DL methods and can be used to validate whether the unsupervised GDD can be comparable to the supervised DL methods.

To qualitatively validate the performance of the methods, the synthetic data (i.e., reduced spatial resolution images) and real data (i.e., original spatial resolution images) were used. Four different criteria were used for evaluation. Similar to the experiments of the HS super-resolution, ERDAS and SA were also considered in pansharpening. In addition, the eight-band extension of average universal image quality index (Q8 [32, 3]), and spatial correlation coefficient (SCC [41]) were used for evaluation. In pansharpening, there are also criteria to validate the performance of the methods on the original spatial resolution images without using reference data. The criteria include a spectral quality index ($\mathbf{D}_\lambda$) and a spatial quality index ($\mathbf{D}_S$), and the joint spectral and spatial quality with no reference (QNR [2]). The criteria were used to validate the methods using real data (i.e., original spatial resolution of images).

Results. Table 2 shows the average results of all test images. When using the synthetic data with reference data, GDD outperformed other existing methods

in terms of all criteria. This showed that GDD reconstructed an HR-MS image that has better quality of both spectral and spatial information. Fig. 7 shows RGB of the reconstructed MS images, the enlarged RGB images, and the corresponding error maps. Although PanNet, PNN, or DRPNN generated sharp edges in the reconstructed images, the spectral information was distorted, which led to the colors that are different from the reference. DIP produced blurred results especially at the edges of the reconstructed images. GDD preserved the spectral information while producing similar spatial details with reference data. This led to lower errors in the reconstructed image. Real images (original resolution images) were also used to evaluate the reconstructed images, as shown in Table 2. DIP produced the lowest value in terms of D_λ. This shows that the spectra reconstructed by DIP are most similar to the spectra of the LR-MS image. PNN+ produced the lowest value in terms of D_s. This shows that the spatial details reconstructed by PNN+ are the most similar to the spatial details of the pansharpening image. GDD performed better than the other methods in terms of QNR. GDD properly balanced the tradeoff between spectral and spatial resolution, which led to the better value of QNR.

Table 2: Average results of different image scenes for pansharpening. Synthetic represents evaluation with reference at lower resolution. Real represents evaluation with no reference at original resolution. ↓ shows lower is better while ↑ shows higher is better.

		BDSD	MTF-GLP	SIRF	DRPNN	PanNet	PNN	PNN+	DIP	GDD
Synthetic	Q8↑	0.8879	0.9074	0.8935	0.9144	0.9164	0.9073	0.9231	0.9171	**0.9469**
	SA↓	5.9425	5.4838	5.9248	5.3690	5.4475	6.5587	5.7963	4.6514	**4.0254**
	ERGAS↓	4.6554	4.1339	3.9836	3.6549	3.9762	4.1547	3.7432	3.5274	**2.6879**
	SCC↑	0.9071	0.9021	0.8970	0.9316	0.8868	0.9131	0.9048	0.8965	**0.9418**
Real	QNR↑	0.9077	0.9157	0.9071	0.8648	0.8833	0.9253	0.9492	0.9446	**0.9517**
	D_λ↓	0.0423	0.0391	0.0538	0.0320	0.0574	0.0316	0.0250	**0.0188**	0.0202
	D_s↓	0.0531	0.0469	0.0414	0.1066	0.0629	0.0447	**0.0264**	0.0374	0.0288

5.3 Denoising

In this section, the reconstruction of a flash image with the help of a no-flash image was addressed to show another application of GDD. The no-flash image acquires an image under ambient illumination where the image can be noisy because of the low-light conditions [23]. However, the flash image acquires an image under artificial light where the image is noise-free and the spatial details of the image are recorded. However, the lighting characteristics are unnatural, and unwanted shadows or artifacts may be produced in the flash image. The objective of this application is to reconstruct a clean no-flash image using the features of a flash image. In this application, true reference data cannot be available.

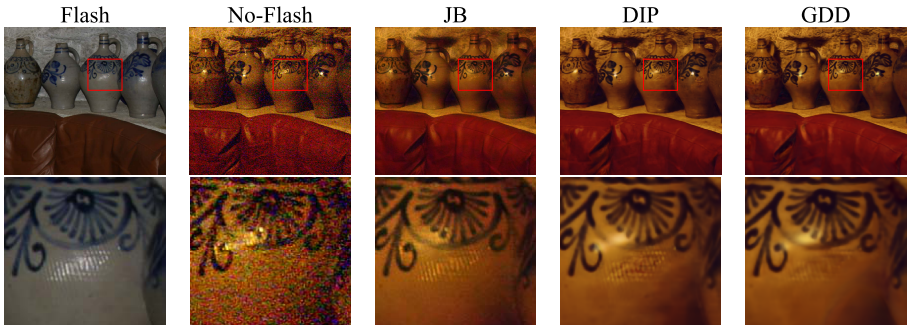

Fig. 8: The reconstructed images of the no-flash image with the help of the flash image.

Although an image with long exposure may be used as a reference [23], the magnitude or characteristics of illumination are not necessarily the same as those of the true reference. In this study, the reconstructed images are qualitatively evaluated according to [28]. In [28], DIP that uses the flash image as an input and the no-flash image as an output was successfully applied to the problem. We qualitatively examined if the architecture used in GDD was as effective as DIP.

Fig. 8 shows that the reconstructed images of the no-flash image. DIP and GDD removed the artifacts more clearly than the joint bilateral method (JB) [23]. GDD produced more explicit boundaries of the image than DIP while preserving the natural colors of the image. This shows that GDD performed at least as well as DIP for the no-flash image reconstruction.

6 Conclusion

We proposed an unsupervised image fusion method that was based on GDD. GDD is a network architecture-based regularizer and can be used to solve different image fusion problems that have been independently studied so far. The network architecture can better exploit spatial details and semantic features of a guidance image. This is achieved by considering an encoder-decoder network that extracts spatial details and semantic features of a guidance image. The multi-scale attention gates enable the extracted semantic features to guide a deep decoder network that generates an output image. This approach achieved the SOTA performance in the different image fusion problems. It pushes the boundaries of the current studies that address only one specific problem. The promising results open up the possibility of a network architecture-based prior that can be used for general purpose including various image fusion problems.

References

1. Akhtar, N., Shafait, F., Mian, A.: Bayesian sparse representation for hyperspectral image super resolution. In: CVPR. pp. 3631–3640 (2015)
2. Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., Selva, M.: Multispectral and panchromatic data fusion assessment without reference. Photogrammetric Engineering and Remote Sensing **74**(2), 193–200 (2008)
3. Alparone, L., Baronti, S., Garzelli, A., Nencini, F.: A global quality measurement of pan-sharpened multispectral imagery. IEEE Geoscience and Remote Sensing Letters **1**(4), 313–317 (2004)
4. Chakrabarty, P., Maji, S.: The spectral bias of the deep image prior. In: NeurIPS Workshops (2019)
5. Chen, C., Li, Y., Liu, W., Huang, Z.: Sirf: Simultaneous satellite image registration and fusion in a unified framework. IEEE Transactions on Image Processing **24**(11), 4213–4224 (2015)
6. Chen, C., Li, Y., Liu, W., Huang, J.: Image fusion with local spectral consistency and dynamic gradient sparsity. In: CVPR (2014)
7. Dian, R., Fang, L., Li, S.: Hyperspectral image super-resolution via non-local sparse tensor factorization. In: CVPR. pp. 3862–3871 (2017)
8. Dian, R., Li, S., Guo, A., Fang, L.: Deep hyperspectral image sharpening. IEEE transactions on neural networks and learning systems **29**(11), 5345–5355 (2018)
9. Dong, W., Fu, F., Shi, G., Cao, X., Wu, J., Li, G., Li, X.: Hyperspectral image super-resolution via non-negative structured sparse representation. IEEE Transactions on Image Processing **25**(5) (2016)
10. Fu, X., Lin, Z., Huang, Y., Ding, X.: A variational pan-sharpening with local gradient constraints. In: CVPR (2019)
11. Fu, Y., Zhang, T., Zheng, Y., Zhang, D., Huang, H.: Hyperspectral image super-resolution with optimized rgb guidance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11661–11670 (2019)
12. Gandelsman, Y., Shocher, A., Irani, M.: "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In: CVPR (June 2019)
13. Garzelli, A., Nencini, F., Capobianco, L.: Optimal mmse pan sharpening of very high resolution multispectral images. IEEE Transactions on Geoscience and Remote Sensing **46**(1), 228–236 (2007)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (June 2016)
15. Heckel, R., Hand, P.: Deep decoder: Concise image representations from untrained non-convolutional networks. In: ICLR (2019)
16. Kawakami, R., Matsushita, Y., Wright, J., Ben-Ezra, M., Tai, Y., Ikeuchi, K.: High-resolution hyperspectral imaging via matrix factorization. In: CVPR. pp. 2329–2336 (2011)
17. Kwon, H., Tai, Y.W.: RGB-guided hyperspectral image upsampling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 307–315 (2015)
18. Lanaras, C., Baltsavias, E., Schindler, K.: Hyperspectral super-resolution by coupled spectral unmixing. In: ICCV (2015)
19. Liu, P., Xiao, L., Li, T.: A variational pan-sharpening method based on spatial fractional-order geometry and spectral–spatial low-rank priors. IEEE Transactions on Geoscience and Remote Sensing **56**, 1788–1802 (2018)
20. Lutio, R.d., D'Aronco, S., Wegner, J.D., Schindler, K.: Guided super-resolution as pixel-to-pixel transformation. In: ICCV (2019)

21. Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G.: Pansharpening by convolutional neural networks. Remote Sensing **8**(7), 594 (2016)
22. Palsson, F., Sveinsson, J.R., Ulfarsson, M.O.: A new pansharpening algorithm based on total variation. IEEE Geoscience and Remote Sensing Letters **11**, 318–322 (2014)
23. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. ACM Transactions on Graphics **23**(3), 664 (2004)
24. Qu, Y., Qi, H., Kwan, C.: Unsupervised sparse dirichlet-net for hyperspectral image super-resolution. In: CVPR (2018)
25. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. vol. 9351, pp. 234–241 (2015)
26. Scarpa, G., Vitale, S., Cozzolino, D.: Target-adaptive cnn-based pansharpening. IEEE Transactions on Geoscience and Remote Sensing **56**(9), 5443–5457 (Sept 2018)
27. Sidorov, O., Hardeberg, J.Y.: Deep hyperspectral prior: Denoising, inpainting, super-resolution. In: ICIP (2019)
28. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: CVPR (2018)
29. Vivone, G., Alparone, L., Chanussot, J., Mura, M.D., Garzelli, A., Licciardi, G., Restaino, R., Wald, L.: A critical comparison among pansharpening algorithms. IEEE Transactions on Geoscience and Remote Sensing **53**(5), 2565–2586 (2014)
30. Wald, L., Ranchin, T., Mangolini, M.: Fusion of satellite images of different spatial resoltuions: assessing the quality of resulting images. Photogrammetric engineering and remote sensing **63**(6), 691–699 (1997)
31. Wald, L.: Quality of high resolution synthesised images: Is there a simple criterion? In: Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images". pp. 99–103. SEE/URISCA (2000)
32. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE signal processing letters **9**(3), 81–84 (2002)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
34. Wei, Q., Dobigeon, N., Tourneret, J., Bioucas-Dias, J., Godsill, S.: R-fuse: Robust fast fusion of multiband images based on solving a sylvester equation. IEEE Signal Processing Letters **23**(11), 1632–1636 (Nov 2016)
35. Wei, Y., Yuan, Q., Shen, H., Zhang, L.: Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geoscience and Remote Sensing Letters **14**(10), 1795–1799 (2017)
36. Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W., Xu, Z.: Multispectral and hyperspectral image fusion by ms/hs fusion net. In: CVPR (2019)
37. Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J.: Pannet: A deep network architecture for pan-sharpening. In: ICCV. pp. 1753–1761 (2017)
38. Yokota, T., Kawai, K., Sakata, M., Kimura, Y., Hontani, H.: Dynamic pet image reconstruction using nonnegative matrix factorization incorporated with deep image prior. In: ICCV (2019)
39. Yokoya, N., Grohnfeldt, C., Chanussot, J.: Hyperspectral and multispectral data fusion: A comparative review of the recent literature. IEEE Geoscience and Remote Sensing Magazine **5**(2), 29–56 (2017)
40. Yokoya, N., Yairi, T., Iwasaki, A.: Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Transactions on Geoscience and Remote Sensing **50**(2), 528–537 (2012)

41. Zhou, J., Civco, D., Silander, J.: A wavelet transform method to merge landsat tm and spot panchromatic data. International journal of remote sensing **19**(4), 743–757 (1998)