A General Paradigm with Detail-Preserving Conditional Invertible Network for Image Fusion

Wu Wang[1] · Liang-Jian Deng[1] · Ran Ran[1] · Gemine Vivone[2,3]

Received: 3 April 2023 / Accepted: 29 September 2023

Abstract

Existing deep learning techniques for image fusion either learn image mapping (LIM) directly, which renders them ineffective at preserving details due to the equal consideration to each pixel, or learn detail mapping (LDM), which only attains a limited level of performance because only details are used for reasoning. The recent lossless invertible network (INN) has demonstrated its detail-preserving ability. However, the direct applicability of INN to the image fusion task is limited by the volume-preserving constraint. Additionally, there is the lack of a consistent detail-preserving image fusion framework to produce satisfactory outcomes. To this aim, we propose a general paradigm for image fusion based on a novel conditional INN (named DCINN). The DCINN paradigm has three core components: a decomposing module that converts image mapping to detail mapping; an auxiliary network (ANet) that extracts auxiliary features directly from source images; and a conditional INN (CINN) that learns the detail mapping based on auxiliary features. The novel design benefits from the advantages of INN, LIM, and LDM approaches while avoiding their disadvantages. Particularly, using INN to LDM can easily meet the volume-preserving constraint while still preserving details. Moreover, since auxiliary features serve as conditional features, the ANet allows for the use of more than just details for reasoning without compromising detail mapping. Extensive experiments on three benchmark fusion problems, i.e., pansharpening, hyperspectral and multispectral image fusion, and infrared and visible image fusion, demonstrate the superiority of our approach compared with recent state-of-the-art methods. The code is available at https://github.com/wwhappylife/DCINN

Keywords Image fusion · Invertible network · Detail preservation · Pansharpening · Hyperspectral and multispectral image fusion · Infrared and visible image fusion · Remote sensing

1 Introduction

Because of physical limitations of imaging devices, there is a trade-off between captured images from different sensors. This trade-off often occurs for details, i.e., spatial details and spectral information trade-off for multispectral pansharpening and hyperspectral and multispectral image fusion (HMF), and salient objects and details trade-off for infrared and visible image fusion (IVF). Thus, the purpose of image fusion is to fuse images from different sources to obtain an image containing complementary information, further achieving the detail-preserving fusion for various tasks, such as pansharpening, HMF, and IVF (considered in this work). A schematic illustration of the different image fusion tasks is shown in Fig. 1.

Compared with traditional image fusion methods (Zhuang et al., 2019; Guo et al., 2020; Yang et al., 2020b, a; Xu et al., 2014; Ma et al., 2016), in the past few years, deep learning

Communicated by Massimiliano Mancini.

✉ Liang-Jian Deng
liangjian.deng@uestc.edu.cn

Wu Wang
wangwu@uestc.edu.cn

Ran Ran
ranran@std.uestc.edu.cn

Gemine Vivone
gemine.vivone@imaa.cnr.it

[1] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China

[2] Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy

[3] NBFC, National Biodiversity Future Center, 90133 Palermo, Italy

Published online: 23 October 2023

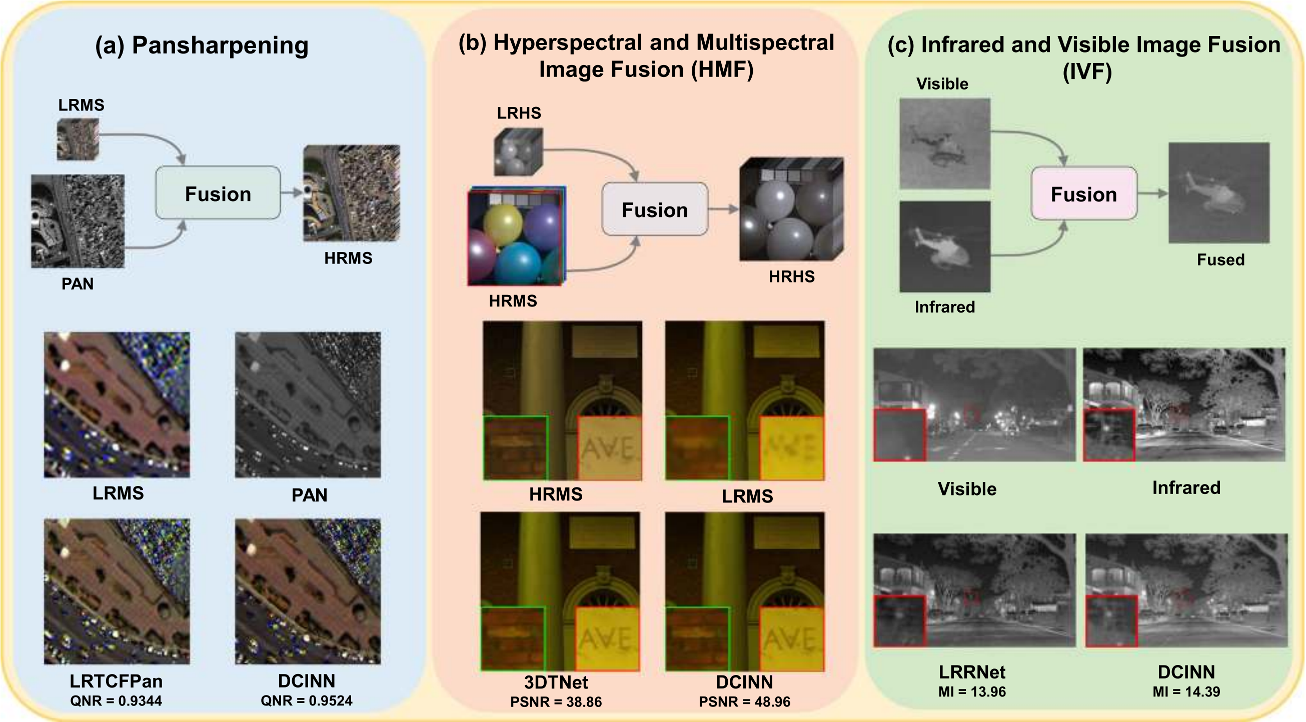


Fig. 1 Schematic illustrations of several image fusion tasks using real examples. First row, flowcharts of three typical image fusion tasks (i.e., pansharpening, HMF, and IVF), where LRMS refers to a low-resolution multispectral image, PAN is a panchromatic image, LRHS indicates a low-resolution hyperspectral image, and so forth. Second and third rows (from left to right): results of the corresponding image fusion tasks by state-of-the-art (SOTA) techniques, such as LRTCFPan (Wu et al., 2023), 3DTNet (Ma et al., 2023), LRRNet (Li et al., 2023), and the proposed DCINN. Compared with prior image fusion methods that can only solve a specific task, the DCINN can be applied to several image fusion tasks producing state-of-the-art outcomes

(DL)-based approaches (Guo et al., 2023; Liu et al., 2023; Ran et al., 2023; Yan et al., 2022) have made significant progress on various image fusion tasks becoming the mainstream framework thanks to their powerful ability to learn a complex mapping from a large number of paired data used for training. These DL-based image fusion methods can be roughly classified into two categories, i.e., learning image mapping (LIM)-based and learning detail mapping (LDM)-based methods.

LIM-based methods (Xu et al., 2022; Tang et al., 2022; Zhou et al., 2023) directly learn the image mapping with DL-based models to preserve the details as shown in Fig. 2a. For example, the method in (Xu et al., 2022) concatenates the source images and uses DenseNet (Huang et al., 2017) to predict the fused image. (Tang et al., 2022) adopts transformer to separately extract features from the source images, then fuse the features to obtain the fused outcome. LIM-based methods directly learn the image mapping, thus getting the advantages of being able to fully utilize the information of source images for reasoning. Their disadvantage is that they do not treat the image intensity discriminating the different frequencies (including the relevant high-frequency details), directly sending the images into the network for training, thus getting a weak feature extraction.

In contrast, as shown in Fig. 2b, the LDM-based methods decompose first the source images into a detail component and a base component, then learning the detail mapping with a residual network and obtaining the outcome by adding the learned details component with the original base component. For example, LPPN (Jin et al., 2022a) considers the low-resolution multispectral (LRMS) image as base component proposing the use of a Laplacian pyramid network to decompose the high-resolution multispectral (HRMS) image into multi-scale details to learn a multi-scale detail mapping. This design has two benefits. First, it allows paying more attention to details. Second, the detail mapping can be learned easier than the image mapping, leading to acceptable results even with a network with reduced capacity. Therefore, the LDM-based methods have been proposed to solve different image fusion tasks. However, their performance is still limited as only details information is taken into account for reasoning.

Whatever (LIM)-based or (LDM)-based image fusion methods, they mainly employ some commonly-used forward-propagation models, such as convolutional neural networks (CNNs) and transformers for feature representation and

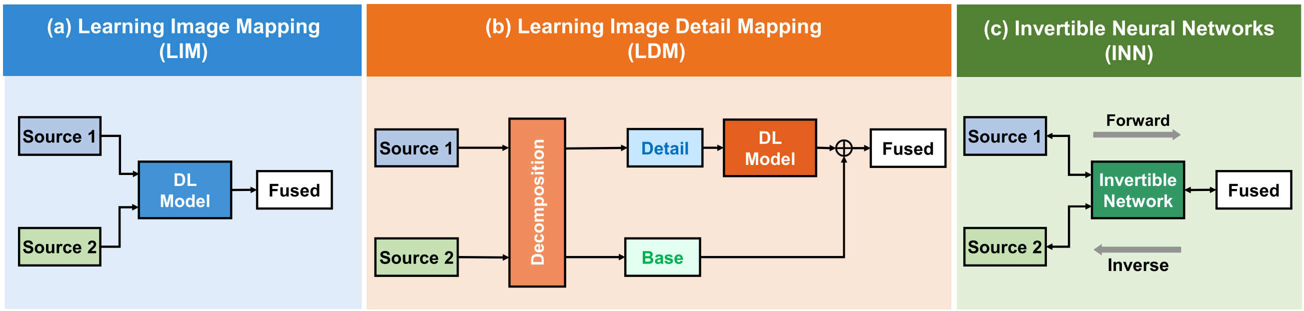


Fig. 2 A comparison of **a** the LIM-based framework, **b** the LDM-based scheme, and **c** the INN paradigm. Because LIM-based methods treat each pixel in a equal way, they are inefficient in preserving details. While LDM-based methods just exploit details (mainly high-frequency local structures) for reasoning. As a result, they are inefficient in capturing low-frequency global structures. Compared with the forward-propagation networks, the INN has a bidirectional structure having no information loss, thus implicitly preserving information. However, the bidirectional structure of INN requires the constraint of volume-preserving, which is generally not met for most (multiple inputs) of the image fusion tasks, motivating us to propose the so-called DCINN approach

extraction, as well as fail to realize the lossless detail-preserving, thus resulting in a poor information learning. Recently, invertible networks (INNs), used in many image applications, such as image resizing (Xiao et al., 2020), image steganography (Lu et al., 2021), and image decolorization (Zhao et al., 2021), have been proposed showing better abilities than the previous commonly-used techniques thanks to their lossless nature. Unfortunately, using INN to directly learn the image mapping for image fusion tasks can encounter two difficulties. The first one is that the invertible ability of INN is originated from its bidirectional structure as shown in Fig. 2c. Therefore, applying INN necessitates volume-preserving, which requires that the total size of the source images is equal to the size of the target image. This constraint is not met in many image fusion tasks. For example, in the pansharpening and HMF tasks, the size of the observations is much smaller than the one of the target image. In the IVF task, the total size of the infrared and visible images is twice of the one of the target image. The second issue is that to realize the invertibility, INNs require severe constraints on network architectures or the related weights, resulting in a weaker capacity than that of the one of commonly-used techniques. A confirmation of this statement is given by the worse classification performance obtained by the INN compared to commonly-used methods exploiting the same level of parameter amount and computation, see (Gomez et al., 2017; Behrmann et al., 2019).

To address the aforementioned issues, we observed that combining INN with LDM-based methods can solve the volume-preserving problem of INN by transforming image mapping into detail mapping. We will show how this design can help in meeting the volume-preserving constraint. Moreover, because details contain less information compared with the whole image, detail mapping is usually easier to learn, thus alleviating the capacity problem of INNs and, in the meantime, exploiting the preservation abilities of INNs for details. Motivated by the advantages and disadvantages of LIM- and LDM-based methods, we propose the use of an auxiliary network (ANet) to extract auxiliary features directly from the source images. The auxiliary features serve as conditional features, and they are fed into the INN to aid the details reasoning. This novel design enables the use of more information for reasoning rather than just details, while maintaining the learned mapping as detail mapping and preserving INN invertibility. The overall framework is denoted as detail-preserving conditional invertible network (DCINN).

We apply the DCINN framework to three widely known image fusion tasks, i.e., pansharpening, HMF, and IVF. In summary, our work has the following contributions:

- We propose a general paradigm based on the DCINN that takes advantage of the benefits of LDM-based, LIM-based, and INN methods while avoiding their drawbacks to more effectively tackle image fusion tasks. To the best of our knowledge, the proposed method is the first paradigm for image fusion that uses a fully invertible network. Further to that, this paradigm can be easily applied to multiple image fusion tasks, such as pansharpening, HMF, and IVF, achieving state-of-the-art results.
- We design the corresponding network modules after carefully analyzing the characteristics of the three image fusion applications, with the goal of utilizing the advantages of LDM-based, LIM-based, and INN methods. Afterwards, we design an auxiliary network and a detail/base decomposition to build a general paradigm that can face with the three fusion tasks. With this paradigm, the fusion procedure becomes fully invertible and overcomes the previously mentioned issues in LDM- and LIM-based techniques.

- Extensive experiments demonstrate that the given DCINN method can achieve state-of-the-art (SOTA) performance on the three above-mentioned fusion tasks. An ablation study verify the need of the novelties introduced in the given paradigm.

2 Related Work

2.1 LDM-Based Methods

For the pansharpening task, to retain both the spectral and spatial information, PanNet (Yang et al., 2017) and DMDNet (Fu et al., 2020) take the LRMS image as base component and extract the detail component from both panchromatic (PAN) and LRMS images using high-pass filters to learn detail mapping. For the HMF task, instead of using predefined filters to extract the detail component, the work in (Xiao et al., 2022) takes the low-resolution hyperspectral (LRHS) image as base component and uses an encoder-decoder network structure to obtain the detail component from the HRMS image. In the IVF community, traditional methods such as (Ma et al., 2016; Adu et al., 2013) customarily decompose the infrared and visible image into base and detail components and fuse each part separately according to some decision rules. Inspired by these traditional works, LDM-based methods (Liu et al., 2020a; Zhao et al., 2020) are also proposed using DL-based models to separately fuse the base and detail components. Nonetheless, these kinds of methods can only use details for reasoning resulting in poor performance.

2.2 LIM-Based Methods

The LIM-based methods mainly focus on the improvement of network architectures, including typical attempts, such as skip-connections (He et al., 2019; Hou et al., 2020), attention mechanisms (Hu et al., 2022b; Guan & Lam, 2021), and transformers (Hu et al., 2022a; Tang et al., 2022). The early LIM-based methods adopt skip connections to train much deeper or wider networks. ResTFNet (Liu et al., 2020b) proposed residual connections to solve the HMF task. DenseFuse (Li & Wu, 2019) introduced dense connections to solve the IVF task. Since the neural networks are known to extract considerable redundant features (Wang et al., 2022), various attention mechanisms have been introduced to allow networks to focus on more representative features. HSRNet (Hu et al., 2022b) proposed spatial-spectral attention to explore the spatial-spectral correspondence of the images for the HMF task. While previous LIM-based methods are mainly based on convolutional networks, which are inefficient in modeling long-range relationships, current LIM-based methods adopt transformers (Deng et al., 2023) to image fusion tasks to address this issue. YDTR (Tang et al., 2022) proposed the use of two-branch transformers to separately extract features from the two source images. The features are added and further fused with transformers to produce the fused outcome to solve the IVF task. More specifically, generative adversarial networks (GANs) have also been proposed to directly learn image mapping for obtaining better fusion products (Ma et al., 2019, 2020b, a). However, the LIM-based methods fail to discriminatively distinguish the base and detail components, thus making them ineffective for preserving details.

2.3 INN

Compared with commonly-used forward propagation CNN structures, INN has a bidirectional structure that simultaneously allows forward and backward propagation, which can theoretically lead to an information lossless model. Due to this favorable property, the INN has received increasing attention in recent years and has been successfully applied in various applications, such as image compression (Xu & Zhang, 2021), image rescaling (Xiao et al., 2020), and image denoising (Huang & Dragotti, 2022). Despite such remarkable progress, the INN has rarely been explored for image fusion tasks. Only a few works have been proposed applying the INN to address the image fusion problem, i.e., Panformer (Zhou et al., 2022) for the pansharpening task, (Cui et al., 2022) and CDDFUse (Zhao et al., 2023) for the IVF task. However, our method has significant differences with them. DCINN is a general paradigm that cannot only solve camera image fusion tasks, i.e., IVF, but can also solve remote sensing image fusion problems, e.g., pansharpening and HMF. Instead, the compared methods can only deal with one of these image fusion problems. Moreover, DCINN is based on the conditional INN while the compared methods are based on the native INN, which may limit the capacity of the INN. Compared with Panformer (Zhou et al., 2022), Panformer utilizes first a transformer branch and a CNN branch to extract features, then enriching these features by a sequential of densely connected invertible modules. Since the combined structure contains noninvertible modules, such as CNN and transformer, the given network is not invertible, which indicates that the benefits of INN cannot be shared by this approach. Compared with CDDFuse (Zhao et al., 2023), CDDFuse uses a pre-trained autoencoder to decompose the source images into base and detail components, then the INN is exploited to preserve details. Instead, our DCINN adopts different strategies to decompose the source images into base and detail components, then using a conditional INN together with an auxiliary network to preserve details. Moreover, our task-specific decomposition strategies enable us to solve HMF and pansharpening tasks, which cannot be solved by CDDFuse (Zhao et al., 2023). The additional auxiliary network of DCINN enables to use more information for reasoning. Compared with IVF-INN (Cui et al., 2022), IVF-INN uses an INN to encode the source images into a latent space and design loss functions to decompose the latent space into the base and detail components, then these components are fused based on some fusion rules and decoded with the same INN to obtain the fused outcome. Instead, DCINN uses conditional INN to learn the detail mapping rather than to encode the source images. Moreover, the decomposition of DCINN is performed on the source images rather than on the latent space. Besides, the basic structures of the INN are also different. More specifically, DCINN additionally uses 5 × 5 invertible convolutional layers (Emiel et al., 2020) to enhance the channel interaction and gated deconvolution feed-forward (GFDN) (Zamir et al., 2022) layers to dynamically fuse information from two streams.

Overall, we think there are two inevitable challenges when applying INN to image fusion tasks in the manner of LIM. Without losing generality, we take a toy example to illustrate the two challenges. Let us suppose that an INN is composed of a sequential of invertible transformations, let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]$ denotes the input, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are two components generated by splitting $\mathbf{x}$ along channel dimension. According to the conventional INN, a simple invertible transformation can be represented as follows:

$$\mathbf{y}_1 = \mathbf{x}_1 + \phi(\mathbf{x}_2), \tag{1}$$

$$\mathbf{y}_2 = \mathbf{x}_2 + \eta(\mathbf{y}_1), \tag{2}$$

where $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2]$ is the output of this transformation, and ϕ and η can be arbitrary neural networks. This transformation can be easily inverted as follows:

$$\mathbf{x}_2 = \mathbf{y}_2 - \eta(\mathbf{y}_1), \tag{3}$$

$$\mathbf{x}_1 = \mathbf{y}_1 - \phi(\mathbf{x}_2), \tag{4}$$

According to the toy example in Eqs. (1)–(4), the invertible transformation ensures that $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2]$ always keeps all of the information of $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]$, which implies that the transformation is theoretically lossless. Therefore, if the INN can be directly applied to image fusion problems, the lossless property ensures that the fused outcome by the INN can preserve all the information of the source images, which is also a main goal of image fusion tasks. However, realizing the invertible transformation requires meeting the volume preservation. Specifically, by the aforementioned toy example,[1] it indicates the following volume (or size) relationships, i.e., $\mathsf{size}(\mathbf{x}_1) = \mathsf{size}(\mathbf{y}_1)$, $\mathsf{size}(\mathbf{x}_2) = \mathsf{size}(\mathbf{y}_2)$,[2] which implies that the total size of the source images must be equal to the total size of the target image (called volume-preserving (Dinh et al., 2015) constraint) when directly applying INN to learn image mapping. However, considering the image fusion tasks having two or more inputs with different sizes, the volume-preserving is generally not valid. For instance, taking pansharpening as example, if an LRMS image with 8 bands has a size of 100 × 100 × 8 and the PAN image has a size of 400 × 400, the total size of the source images (i.e., LRMS and PAN images) is 240, 000 pixels, which is much smaller than that of the target HRMS image (400 × 400 × 8) with total size of 1280, 000 pixels. Therefore, the volume-preserving is not met in the application of pansharpening, leading to the challenge of the fully invertible INN for the fusion task. The same conclusion also exists for the HMF and IVF tasks.

When the total size of the inputs is larger than the total size of the outputs, to apply INN, a general technique to meet the volume-preserving constraint in other compute vision tasks as image compression and image denoising is to have as output of the INN a concatenation of a target image and a redundant image so that the total size of outputs is equal to the total size of the inputs (Xu & Zhang, 2021; Xiao et al., 2020). However, this strategy is not feasible for image fusion tasks for two reasons. The first one is that although some loss functions are proposed to enforce that the redundant image contains no task-related information, the redundant image might still carry useful information. For the image compression and image denoising tasks, their aim is to discard some information, i.e., noise and high-frequency information. Therefore, adopting this strategy is reasonable. On the contrary, image fusion tasks aim to preserve the information of source images, especially details. Adopting this technique to deal with image fusion problems inevitably causes loss of details (we will verify this in Sect. 4.4). The second reason is related to the fact that while this strategy might have the potential to be used to image fusion tasks where the size of the input images is bigger than the size of target images (*i.e.,* VIF task), it cannot be applied to image fusion tasks where the size of the outputs is larger than the size of the inputs (i.e., pansharpening and HMF tasks).

Beyond the above challenge, we also notice that the transformation of conventional INN requires splitting the input into two components along the channel dimension, which can reduce the interaction among channels. As a result, this widely existing challenge of conventional INN can result in a limited capacity of feature representation and extraction compared with other neural networks, such as CNN and transformer.

[1] We just use this toy example to indicate the challenges of applying INN to image fusion. The involved invertible transformation in our CDINN paradigm is much more complex and powerful.

[2] The notation $\mathsf{size}(\mathbf{x})$ denotes the total size (or the so-called volume) of $\mathbf{x}$.

2.4 Motivation

According to the above analysis, LIM-based methods, LDM-based methods, and INN have complementary properties in detail preserving when solving image fusion tasks. This observation motivated us to propose a new general detail-preserving paradigm for image fusion tasks to enjoy the benefits of LIM-based methods, LDM-based methods, and INN while sidestepping their defects. To achieve this goal, we further analyze the flowchart of LDM-based methods and find that the learned mapping is a one-to-one mapping in which the detail component, the base component, and the fused image have the same size. This finding drives us to propose applying INN in an LDM manner, which helps to better preserve the details and avoid the small capacity problem of INN since the detail mapping is much easy to learn. Another advantage is that it could help to meet the constraint of volume-preserving after carefully forcing the size of the detail component in the LDM-based framework to be equal to the size of the residual component (verified in Sect. 3.2). Furthermore, we also seek a way to leverage on more information rather than just the detail component for reasoning (i.e., the weakness of LDM-based methods).

However, utilizing more information for reasoning without breaking the detail paradigm is challenging. To solve this problem, we analyze Eqs. (1)–(4) finding that the neural networks in the transformation are never inverted. Therefore, we can inject additional information into the neural networks without compromising the invertibility. Thus, we design an auxiliary network to extract auxiliary features from both the source images and inject them into the neural networks of the invertible transformation (i.e., the benefits of LDM-based methods). As a result, we can use more than just details for reasoning.

3 Proposed Method

We propose in this paper a uniform framework called DCINN by introducing some special structures, such as auxiliary network and image decomposition, aiming to fully inherit the advantages of LIM-based, LDM-based, and INN approaches. The presented DCINN can distinguish base and detail components, and even realizing the volume-preserving property of INN, thus finally achieving more effective outcomes in an information-lossless manner.

3.1 Overall Flowchart

Figure 3 illustrates an overview of the given DCINN paradigm. From the figure, it is clear that our paradigm basically contains two important parts. One is the image decomposition, which can differ varying the fusion problem, the other is the design of the network architecture that mainly contains a so-called auxiliary network (ANet) and a conditional invertible neural network (CINN) network.

For the image decomposition, we separate first the source images into a base and a detail component according to different fusion tasks. This step serves to achieve two goals: (a) we can learn the detail mapping to better handle the details by using details to predict the residual; (b) it enables meeting the volume-preserving constraint so that an INN can be applied to learn the detail mapping under certain principles. However, learning detail mapping cannot make full use of the information from the source images as only high-frequency information is used for reasoning. To solve this problem, we further design an ANet that directly extracts auxiliary features from source images. Afterwards, the CINN takes these auxiliary features as conditioned information and uses the detail component to predict the residual. The reason why the CINN takes auxiliary features as conditioned information is to keep the invertibility of INN (see Sect. 3.2 for details).

For the network architecture, we mainly design the same network structure, including ANet and CINN, for all the fusion tasks, which demonstrates the robustness of the proposed network (see Sect. 3.3 for details). More in detail, the pansharpening and HMF tasks work in a supervised mode by simulating datasets, while the IVF methods are unsupervised because of the lack of labeled datasets, thus we need to carefully develop the loss functions for the different fusion tasks to ensure effectiveness (see Sect. 3.4 for details).

In what follows, we will present the details of the image decomposition and the network architecture. The loss functions for supervised and unsupervised learning are given in the following sections.

3.2 Image Decomposition

The way to decompose the image plays a crucial role in our fusion paradigm. It requires us to separate the base and detail components from the source images in a simple and reasonable manner, aiming to fit the given uniform paradigm. For simplicity, we denote $\mathbf{I}_1$ and $\mathbf{I}_2$ as the two source images.[3], $\widehat{\mathbf{I}}_\mathrm{f}$ as the fused image, $\mathbf{I}_\mathrm{f}$ as the reference image, and $\mathbf{I}_\mathrm{r} = \mathbf{I}_\mathrm{f} - \mathbf{I}_\mathrm{b}$ as the residual image. We decompose the source images into a detail component, $\mathbf{I}_\mathrm{d}$, and a base component, $\mathbf{I}_\mathrm{b}$, so that we can use $\mathbf{I}_\mathrm{d}$ to predict the residual image $\mathbf{I}_\mathrm{r}$ conditioned on $\mathbf{I}_1$ and $\mathbf{I}_2$. We design the decomposition methods under the principle of $\mathsf{size}(\mathbf{I}_\mathrm{b}) = \mathsf{size}(\mathbf{I}_\mathrm{d}) = \mathsf{size}(\mathbf{I}_\mathrm{f})$ to meet the constraint of volume-preserving, such that an INN is utilized to learn the detail mapping to better preserve the details. Note

[3] For the different tasks, the two source images are different. For example, the two source images are the PAN image and LRMS image for pansharpening. More details about the other applications can be found from the bottom part of Fig. 3

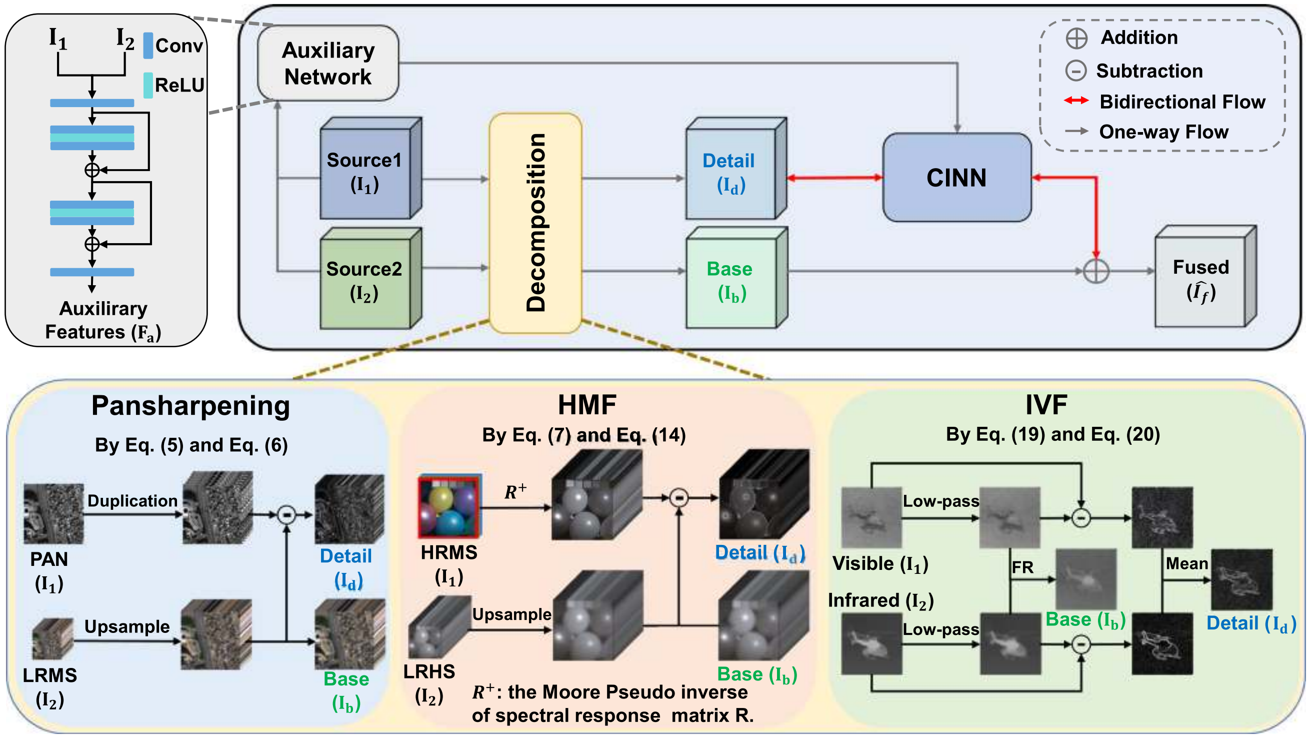


Fig. 3 The flowchart of our DCINN paradigm. The decomposition module transforms the source images into base and detail components. As a result, the volume-preserving is met, allowing us to apply INN to image fusion tasks to better preserve details. The auxiliary network feeds conditional information to the CINN, allowing it to use more than just details for reasoning, while still learning detail mapping and retaining invertibility. The FR in the IVF application indicates the fusion rule in Eq. (19)

that more details about the used notations can be found in Table 1.

Table 1 The involved notations and the corresponding description

Notation	Description
$\mathbf{I}_\mathrm{b}$	Base component
$\mathbf{I}_\mathrm{d}$	Detail component
$\mathbf{I}_\mathrm{r}$	Residual image
$\widehat{\mathbf{I}}_\mathrm{f}$	Fused image
$\mathbf{I}_\mathrm{f}$	Ground-truth (GT) image (reference image)
$\mathbf{F}_\mathrm{a}$	Auxiliary feature
$\mathbf{L}$	Low-resolution multispectral image
$\mathbf{P}$	Panchromatic image
$\mathbf{Y}$	Low-resolution hyperspectral image
$\mathbf{Z}$	High-resolution multispectral image
$\mathbf{X}$	High-resolution hyperspectral image
$\mathbf{B}$	Blur matrix
$\mathbf{R}$	Spectral response matrix
$\mathbf{I}^\mathrm{vi}$	Visible image
$\mathbf{I}^\mathrm{ir}$	Infrared image
FR	Fusion rule

3.2.1 Decomposition for Pansharpening

Let $\mathbf{L} \in \mathbb{R}^{C\times n}$ denotes the LRMS image,[4] $\mathbf{P} \in \mathbb{R}^{1\times N}$ denotes the PAN image, and $\mathbf{H} \in \mathbb{R}^{C\times N}$ denotes the target HRMS image, where C is the number of bands and N is the number of pixels. $\mathbf{L}$ is upsampled to reach the same resolution as $\mathbf{P}$ (Xu et al., 2014). Let $\widehat{\mathbf{L}} \in \mathbb{R}^{C\times N}$ denote the upsampled LRMS image. The goal of the pansharpening task is to generate a fused image that possesses both the spectral information of $\widehat{\mathbf{L}}$ and the spatial information of $\mathbf{P}$. From previous LDM-based methods, e.g., (Deng et al., 2021), $\widehat{\mathbf{L}}$ is considered as base component:

$$\mathbf{I}_\mathrm{b} \triangleq \widehat{\mathbf{L}}, \tag{5}$$

where the symbol "$\triangleq$" means defined as. Then, the detail component $\mathbf{I}_\mathrm{d}$ is extracted by subtracting $\mathbf{I}_\mathrm{b}$ from the replicated $\mathbf{P}$:

$$\mathbf{I}_\mathrm{d} \triangleq \widehat{\mathbf{P}} - \mathbf{I}_b, \tag{6}$$

[4] For convenience, we unfold the two-dimensional images into one-dimensional vectors. The same for the subsequent notations.

where $\widehat{\mathbf{P}} \in \mathbb{R}^{C\times N}$ denotes the replicated version of $\mathbf{P}$. This work focuses on CINN that uses $\mathbf{I}_\mathrm{d} \in \mathbb{R}^{C\times N}$ to predict the residual image ($\mathbf{I}_\mathrm{r} = \mathbf{H} - \mathbf{I}_\mathrm{b} \in \mathbb{R}^{C\times N}$) conditioned on the auxiliary features ($\mathbf{F}_\mathrm{a} = \mathrm{ANet}(\widehat{\mathbf{L}}, \widehat{\mathbf{P}})$). It is clear that the size of $\mathbf{I}_\mathrm{d}$ is equal to the size of $\mathbf{I}_\mathrm{f}$ and $\mathbf{I}_\mathrm{r}$, which indicates that the volume-preserving is satisfied.

3.2.2 Decomposition for HMF

Let $\mathbf{X} \in \mathbb{R}^{C\times N}$ denote the high-resolution hyperspectral (HRHS) image, $\mathbf{Y} \in \mathbb{R}^{C\times n}$ indicates the low-resolution hyperspectral (LRHS) image, and $\mathbf{Z} \in \mathbb{R}^{c\times N}$ is the HRMS image. The target of HMF is to increase the spatial information of the LRHS image, $\mathbf{Y}$, with the HRMS image, $\mathbf{Z}$, while preserving its spectral information. For simplicity, we interpolate, $\mathbf{Y}$, to have the same spatial resolution as $\mathbf{Z}$ and denote it as the upsampled LRHS image, $\widehat{\mathbf{Y}} \in \mathbb{R}^{C\times N}$. Since $\widehat{\mathbf{Y}}$ is spatially degenerated from $\mathbf{X}$, we consider $\widehat{\mathbf{Y}}$ to be the base component:

$$\mathbf{I}_\mathrm{b} \triangleq \widehat{\mathbf{Y}}. \tag{7}$$

After defining the base component, we should extract the details from $\mathbf{Z}$ to define $\mathbf{I}_\mathrm{d}$. Because hyperspectral images have large spectral variances, the replication of $\mathbf{Z}$ computing residuals by $\widehat{\mathbf{Y}}$ to obtain $\mathbf{I}_\mathrm{d}$ can lead to inaccurate results. Thus, we rely on the observation model to precisely extract the detail component. The observation model of the HMF task can be expressed as:

$$\mathbf{Y} = \mathbf{XBS}, \tag{8}$$

$$\mathbf{Z} = \mathbf{RX}, \tag{9}$$

where $\mathbf{R} \in \mathbb{R}^{c\times C}$ is the spectral response matrix, $\mathbf{B} \in \mathbb{R}^{N\times N}$ denotes the blurry matrix, and $\mathbf{S} \in \mathbb{R}^{N\times n}$ stands for the downsampling matrix. For the linear relationship expressed in Eq. (9), we introduce a widely-used concept in matrix theory, i.e., the Moore-Pseudo inverse. We denote $\mathbf{R}^+ \in \mathbb{R}^{C\times c}$ as the Moore-Pseudo inverse of $\mathbf{R}$[5]. Then, it is clear to have the relation $\mathbf{RR}^+\mathbf{R} = \mathbf{R}$ (Barata & Hussein, 2012). According to the observation model in Eq. (9), the matrix $\mathbf{R}^+\mathbf{Z}$ can be further regarded as an approximation of $\mathbf{X}$. Again, it is easy to check that $\mathbf{R}^+\mathbf{Z}$ theoretically preserves all the information of $\mathbf{Z}$ since $\mathbf{R}^+\mathbf{Z}$ is an invertible transformation of $\mathbf{Z}$, which is guaranteed by the following Proposition 1.

Proposition 1 *$\mathbf{R}^+\mathbf{Z}$ is an invertible transformation of $\mathbf{Z}$ and thus preserves all the information of $\mathbf{Z}$.*

Proof By introducing Moore-Pseudo inverse in matrix theory, it is easy to have the following equation with $\mathbf{Z} = \mathbf{RX}$:

$$\mathbf{RR}^+\mathbf{Z} = \mathbf{RR}^+(\mathbf{RX}) = (\mathbf{RR}^+\mathbf{R})\mathbf{X}. \tag{10}$$

According to the property of the Moore-Pseudo inverse (Barata & Hussein, 2012), we have $\mathbf{RR}^+\mathbf{R} = \mathbf{R}$, thus:

$$(\mathbf{RR}^+\mathbf{R})\mathbf{X} = \mathbf{RX} = \mathbf{Z}. \tag{11}$$

Using Eqs. (10)–(11), we have:

$$\mathbf{RR}^+\mathbf{Z} = \mathbf{Z}, \tag{12}$$

which indicates that: (1) $\mathbf{R}^+\mathbf{Z}$ is an invertible transformation of $\mathbf{Z}$ since it can fully recover $\mathbf{Z}$ after the multiplication of $\mathbf{R}^+$ and $\mathbf{R}$; (2) $\mathbf{R}^+\mathbf{Z}$ preserves all the information of $\mathbf{Z}$, since if $\mathbf{R}^+\mathbf{Z}$ fails to hold all the information of $\mathbf{Z}$, the invertible transformation $\mathbf{R}$ cannot recover $\mathbf{R}^+\mathbf{Z}$ to get $\mathbf{Z}$. Note that the analysis of the invertible property is a crucial point in the design of INN from the theoretical perspective.

To derive the detail component, $\mathbf{I}_\mathrm{d}$, for HMF, we have that it is equal to the residual image, $\mathbf{I}_\mathrm{r}$, i.e.:

$$\mathbf{I}_\mathrm{d} = \mathbf{I}_\mathrm{r} = \mathbf{X} - \mathbf{I}_\mathrm{b}, \tag{13}$$

where $\mathbf{I}_\mathrm{b}$ is the upsampled LRHS image, $\widehat{\mathbf{Y}}$, that mainly contains low-frequency information (discussed at the beginning of Sect. 3.2.2), and $\mathbf{X}$ is the latent HRHS. However, due to the unavailable of $\mathbf{X}$, we alternatively take the invertible $\mathbf{R}^+\mathbf{Z}$, whose invertibility has been clearly illustrated, to approximately represent $\mathbf{X}$. $\mathbf{R}^+\mathbf{Z}$ can effectively increase the spectral dimension, in the meanwhile holding all the information of $\mathbf{Z}$ according to Proposition 1. Therefore, we have the following alternative formula to calculate the residual image $\mathbf{I}_\mathrm{d}$:

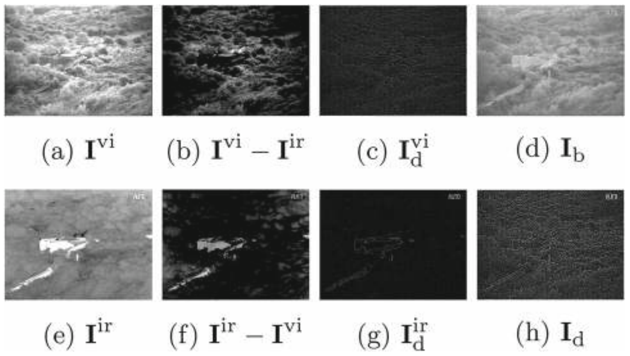
$$\mathbf{I}_\mathrm{d} \triangleq \mathbf{R}^+\mathbf{Z} - \mathbf{I}_\mathrm{b}. \tag{14}$$

Again, the size of $\mathbf{I}_\mathrm{d}$ is equal to the size of $\mathbf{I}_\mathrm{r}$, which still meets the volume-preserving constraint. It is worth noting that the invertibility of $\mathbf{R}^+\mathbf{Z}$ is very important as it ensures that the details of $\mathbf{Z}$ can be preserved after decomposition, which is also mentioned after Eq. (12).

3.2.3 Decomposition for IVF

Let $\mathbf{I}^\mathrm{ir} \in \mathbb{R}^{1\times N}$ be the infrared (IR) image and let $\mathbf{I}^\mathrm{vi} \in \mathbb{R}^{1\times N}$ stand for the visible (VI) image. In general, IVF aims to fuse the salient objects of $\mathbf{I}^\mathrm{ir}$ and the background details of $\mathbf{I}^\mathrm{vi}$ to get an image with clear objects and details. Note that the IVF task has three major differences from both HMF and pansharpening tasks. First, the two source images $\mathbf{I}^\mathrm{vi}$ and

[5] $\mathbf{R}^+$ can be easily computed by using the Matlab function “pinv(R)”.



(a) $\mathbf{I}^{\mathrm{vi}}$ (b) $\mathbf{I}^{\mathrm{vi}} - \mathbf{I}^{\mathrm{ir}}$ (c) $\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}$ (d) $\mathbf{I}_{\mathrm{b}}$

(e) $\mathbf{I}^{\mathrm{ir}}$ (f) $\mathbf{I}^{\mathrm{ir}} - \mathbf{I}^{\mathrm{vi}}$ (g) $\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}}$ (h) $\mathbf{I}_{\mathrm{d}}$

Fig. 4 **IR** and **VI** images and their corresponding details for the application of IVF. **a** and **e** are the visible image and infrared image, respectively. **c** and **g** are the detail components of the visible and infrared images, respectively. The base component in (**d**) and the detail component in (**h**) are obtained from Eqs. (19) and (20), respectively

$\mathbf{I}^{\mathrm{ir}}$ in the IVF both contain details, whereas the details in the HMF and pansharpening tasks are only retained in one of two source images. Second, $\mathbf{I}^{\mathrm{ir}}$ and $\mathbf{I}^{\mathrm{vi}}$ have very different distributions as they are from different modalities. Therefore, the strategy of extracting details by computing the residual image between the source images is unfeasible. Besides, the reference images are not available in the application of IVF since we do not have reasonable ways to simulate datasets. Thus, the training of the network is done in an unsupervised way. As presented in Fig. 4a and e, $\mathbf{I}^{\mathrm{vi}}$ and $\mathbf{I}^{\mathrm{ir}}$ show significantly different appearances. Moreover, their residual images, i.e., Fig. 4b and f, contain both base and detail components. Therefore, following the idea of traditional methods, such as (Ma et al., 2016; Zhou et al., 2016), we decompose first one of the source images, such as $\mathbf{I}^{\mathrm{vi}}$ into one base component (i.e., Eq. 15) and one detail component (i.e., Eq. 17). Similarly, we do the same operation with the other source image, $\mathbf{I}^{\mathrm{ir}}$, to get another pair of base and detail components, i.e., one base component (see Eq. 16) and one detail component (i.e., Eq. 18). To obtain the aforementioned base and detail components, we need to apply first a low-pass (LP) filter[6] to both $\mathbf{I}^{\mathrm{ir}}$ and $\mathbf{I}^{\mathrm{vi}}$ to produce their corresponding base components (i.e.,$\mathbf{I}_{\mathrm{b}}^{\mathrm{ir}}$ and $\mathbf{I}_{\mathrm{b}}^{\mathrm{vi}}$):

$$\mathbf{I}_{\mathrm{b}}^{\mathrm{ir}} = LP\left(\mathbf{I}^{\mathrm{ir}}\right), \tag{15}$$

$$\mathbf{I}_{\mathrm{b}}^{\mathrm{vi}} = LP\left(\mathbf{I}^{\mathrm{vi}}\right). \tag{16}$$

After getting the above base components, then the corresponding detail components (i.e., $\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}}$ and $\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}$) can be calculated as follows:

$$\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}} = \mathbf{I}^{\mathrm{ir}} - \mathbf{I}_{\mathrm{b}}^{\mathrm{ir}}, \tag{17}$$

$$\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}} = \mathbf{I}^{\mathrm{vi}} - \mathbf{I}_{\mathrm{b}}^{\mathrm{vi}}. \tag{18}$$

By checking Fig. 4c and g, it is clear that $\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}}$ and $\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}$ contain complementary information, which can be utilized for the final data fusion step. Based on the aforementioned base and detail components by Eqs. (15)–(18), we can simply define the base component, $\mathbf{I}_{\mathrm{b}}$ (see Fig. 4d), as a fusion product of two base components using a fusion rule:

$$\mathbf{I}_{\mathrm{b}} \triangleq \mathrm{FR}\left(\mathbf{I}_{\mathrm{b}}^{\mathrm{ir}} + \mathbf{I}_{\mathrm{b}}^{\mathrm{vi}}\right), \tag{19}$$

where FR indicates the fusion rule, which could be a weighted average, a max operation, or any other fusion algorithm such as IFCNN (Zhang et al., 2020). It is worth noting that there exist various fusion methods to obtain base components, and these methods are compatible with our method. After defining the base component, $\mathbf{I}_{\mathrm{b}}$, we also use a similar strategy as in Eq. (19) to averagely generate the detail component, $\mathbf{I}_{\mathrm{d}}$ (see Fig. 4h) with the consideration of preserving the information of both $\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}}$ and $\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}$:

$$\mathbf{I}_{\mathrm{d}} \triangleq \left(\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}} + \mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}\right)/2. \tag{20}$$

In summary, we obtained the base component, $\mathbf{I}_{\mathrm{b}}$, and the detail component, $\mathbf{I}_{\mathrm{d}}$, for the three fusion tasks, i.e., pansharpening, HMF, and IVF. The obtained components can be included into our general framework for the final fusion step. For simplicity, we list these components for each task at the bottom of Fig. 3.

3.3 Network Architecture

In this section, we will introduce the network architecture involved in the proposed approach, including some innovative network designs that can effectively embed the base and detail components decomposed in the previous section within a unified methodology framework, as well as enabling information lossless by ensuring the volume-preserving of INN. The overall network architecture is presented in Fig. 3.

Having a look at Fig. 3, it is clear that our paradigm mainly involves a CINN and an ANet that aims to learn an invertible detail mapping with sufficient information of source images to better preserve details. Specifically, the ANet takes in input two source images to generate the auxiliary features $\mathbf{F}_{\mathrm{a}}$, then the CINN considers $\mathbf{F}_{\mathrm{a}}$ and $\mathbf{I}_{\mathrm{d}}$ to predict $\mathbf{I}_{\mathrm{r}}$. Note that directly concatenating $\mathbf{F}_{\mathrm{a}}$ and $\mathbf{I}_{\mathrm{d}}$ and sending them to CINN leads to the volume-preserving constraint that cannot be valid, thus the learned mapping is not invertible. Instead, the CINN gets $\mathbf{F}_{\mathrm{a}}$ as conditional information and uses $\mathbf{I}_{\mathrm{d}}$ to predict $\mathbf{I}_{\mathrm{r}}$, which can meet the volume-preserving constraint. In what follows, we will introduce in detail the structures of the CINN and ANet. We will also point out how to build the CINN such

[6] In the experiments, the low-pass filter is a zero-mean Gaussian filter with size of 11 × 11 and standard deviation equal to 1.

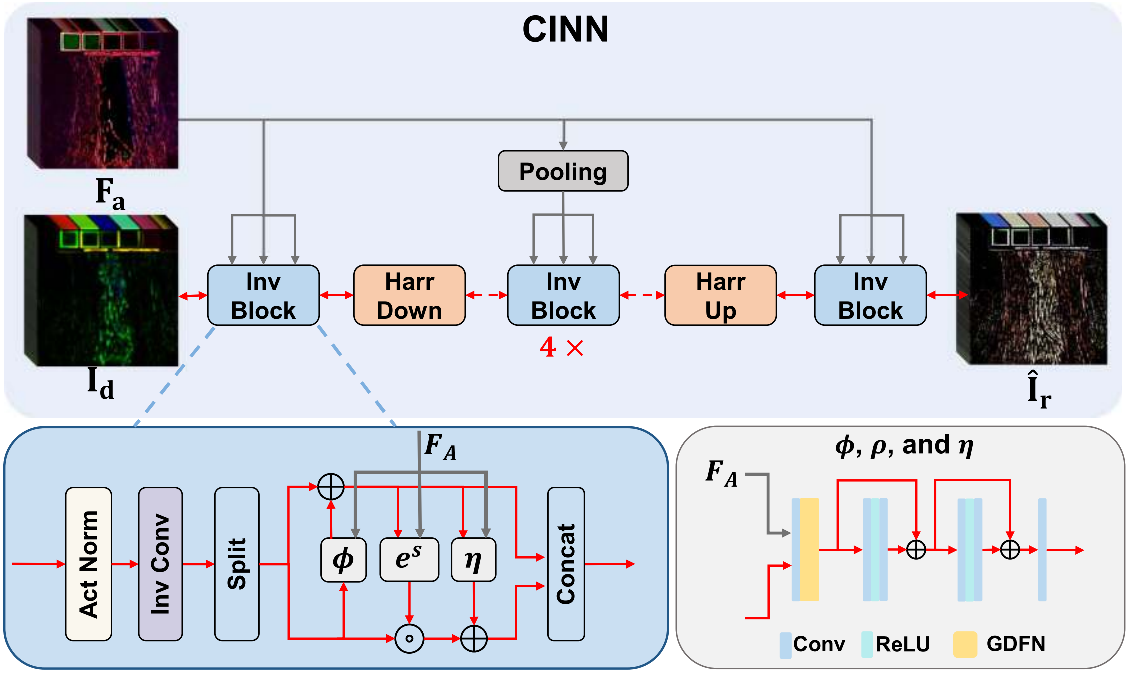


Fig. 5 The structure of the proposed CINN. CINN predicts the residual image using the detail component based on the auxiliary features. The red arrows represent the flow of the details, whereas the black arrows represent the flow of $\mathbf{F}_\mathrm{a}$. It is made up of invertible blocks (Inv Blocks) for feature extraction and the Haar transform (Harr Down and Harr Up) for feature resizing. The general structure of CINN is shown at the top. Bottom left: the structure of an invertible block, where $\phi(\cdot)$, $s(\cdot)$, and $\eta(\cdot)$ are neural network transformations (see Eqs. (23)-(25)). Bottom right: detailed structures of $\phi(\cdot)$, $\rho(\cdot)$, and $\eta(\cdot)$

that it can use $\mathbf{F}_\mathrm{a}$ as conditional information to favor the reasoning while holding the volume-preserving constraint.

3.3.1 The Structure of the Proposed CINN

For the tasks of pansharpening and HMF, Fig. 5 shows the structure of the proposed CINN. Our CINN mainly consists of two invertible blocks that work at the original scale and four invertible blocks that work at a lower (with a factor of 2) scale. This two-scale design enables the better extraction of multi-scale features while maintaining a good balance of computation and parameter efficiency. Specifically, we employ the classical Haar transform (Ardizzone et al., 2019) to downsample and upsample features, since the Haar transform is simple and invertible, only causing limited distortion.

For the IVF task, since both the source images $\mathbf{I}^{\mathrm{ir}} \in \mathbb{R}^{1\times N}$ and $\mathbf{I}^{\mathrm{vi}} \in \mathbb{R}^{1\times N}$ have only one channel, $\mathbf{I}_\mathrm{d} \in \mathbb{R}^{1\times N}$ and $\mathbf{I}_\mathrm{b} \in \mathbb{R}^{1\times N}$ obtained by our decomposition module also have only one channel. As shown in the lower part of Fig. 5, the invertible block needs to split the input features or images along the channel to achieve invertibility, but both $\mathbf{I}_\mathrm{d}$ and $\mathbf{I}_\mathrm{b}$ are single channels and cannot be split. Therefore, it is not feasible to directly use the same CINN as for pansharpening and HMF. For this reason, before sending $\mathbf{I}_\mathrm{d}$ into the CINN, we use first the Harr transform to downsample $\mathbf{I}_\mathrm{d}$ by a factor of 2 to increase the number of channels.[7] The downsampled $\mathbf{I}_\mathrm{d}$ has a size of $4 \times N/4$. Then, CINN uses the downsampled $\mathbf{I}_\mathrm{d}$ and $\mathbf{F}_\mathrm{a}$ to generate a low-resolution $\widehat{\mathbf{I}}_\mathrm{r}$, which also has a size of $4 \times N/4$. Finally, we use another Harr transform to upsample the low-resolution $\widehat{\mathbf{I}}_\mathrm{r}$ by a factor of 2 to reach the original resolution.

As in Fig. 5, several invertible blocks of our CINN have fed by the detail features, $\mathbf{I}_\mathrm{d}$, obtained by the decomposition in Sect. 3.2, and the auxiliary features, $\mathbf{F}_\mathrm{a}$, generated by the ANet (see Sect. 3.3.2). In the invertible blocks, the detail features are normalized first by ActNorm (Kingma & Dhariwal, 2018) for a more stable convergence, subsequently, the normalized features are processed by an invertible convolution with size of 5×5 (Emiel et al., 2020). Compared with an invertible convolution with size of 1×1 in the previous INN, the invertible convolution with size of 5×5 not only enhances the channel interaction,[8] but also improves spatial interaction. The invertible convolution with size of 5×5 can be built using the following matrix exponential with the man-

[7] The Harr transform is an invertible transform that satisfies the volume-preserving constraint by increasing the number of channels when downsampling the image.

[8] We mentioned that INN has a small capacity due to feature splitting. Thus, enhancing the channel interaction is very important for INN to increase the capacity.

ner of Taylor expansion (please, refer to (Emiel et al., 2020) for more details):

$$\exp(\mathbf{M}) = \mathbf{I} + \frac{\mathbf{M}}{1!} + \frac{\mathbf{M}^2}{2!} + \cdots = \sum_{i=0}^{\infty} \frac{\mathbf{M}^i}{i!}, \tag{21}$$

where $\mathbf{M}$ is the square weight matrix of a plain 5×5 convolution and $\mathbf{I}$ is the identity matrix. Accordingly, the inverse transformation is easy to compute as follows:

$$\exp(\mathbf{M})^{-1} = \exp(-\mathbf{M}). \tag{22}$$

Moreover, the detail features are split into two parts, equally along channels. Subsequently, the split detail features together with the auxiliary features are then processed by a specially designed layer with the following affine coupling transformation:

$$\mathbf{y}_1 = \mathbf{x}_1 + \phi(\mathbf{x}_2, \mathbf{F}_\mathrm{a}), \tag{23}$$

$$\mathbf{y}_2 = \mathbf{x}_2 \odot e^{\mathbf{s}} + \eta(\mathbf{y}_1, \mathbf{F}_\mathrm{a}), \tag{24}$$

where $\odot$ represents the element-wise multiplication, $\mathbf{x}_1, \mathbf{x}_2$ are the split features, $\mathbf{y}_1, \mathbf{y}_2$ are the corresponding outputs of the transformation, ϕ and η are two learnable neural networks, e is the natural constant, and $\mathbf{s}$ is a learnable scaling factor (Xiao et al., 2020) with the same size of $\mathbf{x}_2$, which can be computed as follows:

$$\mathbf{s} = 2\sigma(\rho(\mathbf{y}_1, \mathbf{F}_\mathrm{a})) - 1, \tag{25}$$

where σ denotes the sigmoid function, and ρ is another learnable neural network. From the bottom-right part of Fig. 5, we can find the details of the involved three neural networks, i.e., ϕ, ρ, and η, which have the same structure. Specifically, these neural networks all have six convolution layers and a GFDN layer whose details can be referred to (Zamir et al., 2022). Besides, the gating mechanism of GFDN can dynamically fuse the features of the detail component (i.e., $\mathbf{I}_\mathrm{d}$) and the output of the ANet (i.e., $\mathbf{F}_\mathrm{a}$).

More in detail, the inverse of the above affine coupling layer (i.e., Eq. 23) is given by:

$$\mathbf{x}_2 = (\mathbf{y}_2 - \eta(\mathbf{y}_1, \mathbf{F}_\mathrm{a})) \odot e^{-\mathbf{s}}, \tag{26}$$

$$\mathbf{x}_1 = \mathbf{y}_1 - \phi(\mathbf{x}_2, \mathbf{F}_\mathrm{a}). \tag{27}$$

As we can see from the above equations, $\mathbf{F}_\mathrm{a}$ is viewed as conditional information because it is used in both forward and inverse computing. Compared to invertible blocks used in other tasks (Xiao et al., 2020; Huang & Dragotti, 2022; Xu & Zhang, 2021), the proposed invertible block has three differences: (a) our invertible block introduces conditional information to assist reasoning while maintaining invertibility, thereby achieving better performance; (b) we additionally introduce the GFDN (Zamir et al., 2022) to fuse the detail features and auxiliary features; (c) we apply an invertible convolution with a size of 5×5 (Emiel et al., 2020) to enhance the spatial and channel interaction rather than using the invertible convolution with a size of 1×1.

3.3.2 The Design of the Proposed ANet

As mentioned before, the purpose of the auxiliary network (ANet) is to assist the CINN by utilizing all the information (not just details) from the source images. The top-left part of Fig. 3 shows the structure of the ANet, which mainly contains a convolution layer and two residual blocks. Note that the ANet and the CINN will be trained together for better learning. In Fig. 6, we show the learned features with and without the auxiliary network. As we have seen, since the CINN could effectively learn the detail mapping, the extracted features are mainly with high-frequency local structures, while the features learned by the ANet are with more low-frequency global structures. Further, we compared the features shown in the first and second rows of Fig. 6, respectively, it is easy to know that the features in the first row have sharper edges, richer textures, and less activation on the smooth area. By this

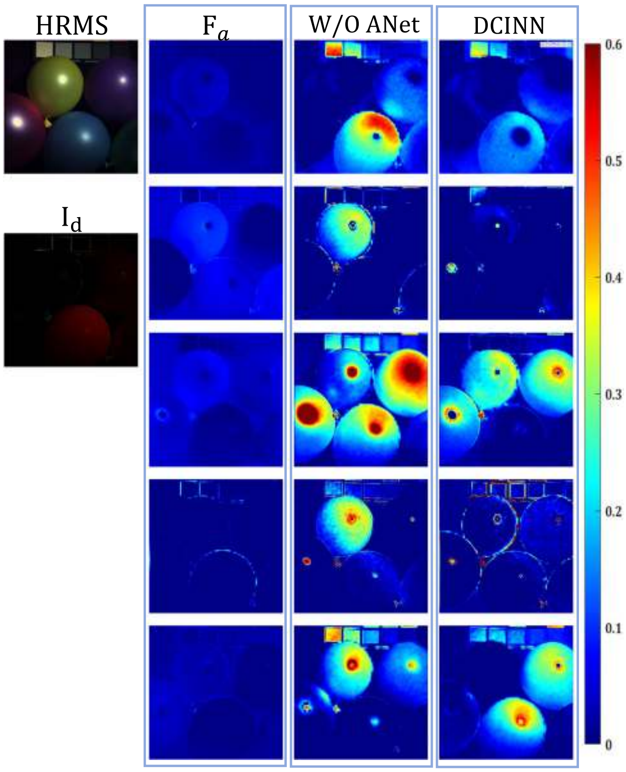


Fig. 6 Visual comparisons of the extracted features of the first invertible block of our DCINN trained on the CAVE dataset without (third column) and with (fourth column) the auxiliary network. The first column is the HRMS image (top) and the corresponding $\mathbf{I}_\mathrm{d}$ (bottom). The second column shows the auxiliary features $\mathbf{F}_\mathrm{a}$ of the ANet

phenomenon (also verified in the ablation study and Fig. 15 in the experiment section), we may conclude that the ANet could extract global structures to guide the learning of local structures.

3.4 Loss Function

Since DCINN relies upon a bidirectional computation, the loss function used for all the involved image fusion problems consists of two items: a forward loss function that enforces the learned residual image, $\widehat{\mathbf{I}}_{\mathrm{r}}$, to be consistent with the ground-truth (GT) residual image, $\mathbf{I}_{\mathrm{r}}$, and a backward loss function that enforces the reconstructed detail component to be consistent with the extracted detail component. Note that the proposed backward loss can favor DCINN in exploring unseen datasets (see Sects. 4.1.4 and 4.2.4). The forward loss function is:

$$\mathcal{L}_{\mathrm{f}} = \ell\left(\widehat{\mathbf{I}}_{\mathrm{r}},\ \mathbf{I}_{\mathrm{r}}\right), \tag{28}$$

where $\widehat{\mathbf{I}}_{\mathrm{r}}$ is the learned residual image, $\mathbf{I}_{\mathrm{r}}$ is the known reference residual image, and ℓ is a function that measures the consistency between $\widehat{\mathbf{I}}_{\mathrm{r}}$ and $\mathbf{I}_{\mathrm{r}}$. Besides, the backward loss function is:

$$\mathcal{L}_{\mathrm{b}} = \ell(\widehat{\mathbf{I}}_{\mathrm{d}},\ \mathbf{I}_{\mathrm{d}}), \tag{29}$$

where $\widehat{\mathbf{I}}_{\mathrm{d}}$ is the reconstructed detail component by the inverse computing of CINN, i.e., $\widehat{\mathbf{I}}_{\mathrm{d}} = \mathrm{CINN}^{-1}(\widehat{\mathbf{I}}_{\mathrm{r}}|\mathbf{F}_{\mathrm{a}})$[9]. The overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{f}} + \lambda\mathcal{L}_{\mathrm{b}}, \tag{30}$$

combining the forward and backward losses through a positive balance parameter λ.

This work mainly involves three image fusion tasks, including two supervised applications (pansharpening and HMF) and one unsupervised application (IVF).[10] Thus, we have to design different $\mathcal{L}_{\mathrm{f}}$ and $\mathcal{L}_{\mathrm{b}}$ for the two types of tasks.

For the supervised task, we only utilize the conventional ℓ_1 norm for both $\mathcal{L}_{\mathrm{f}}$ and $\mathcal{L}_{\mathrm{b}}$ as it has been proved to have the capacity of detail-preserving.

For the unsupervised task, i.e., IVF, we empirically find that the backward loss function has no performance benefits. Therefore, we set λ to 0. For the forward loss function, we refer to (Tang et al., 2022) using the structural similarity index measure (SSIM) (Wang et al., 2003) loss and the gradient-based loss as forward loss function. The SSIM (Wang et al., 2003) loss function is given as follows:

$$\mathcal{L}_{\mathrm{s}} = \left(1 - \mathrm{SSIM}\left(\widehat{\mathbf{I}}_{\mathrm{f}},\ \mathbf{I}^{\mathrm{ir}}\right)\right) + \beta_1\left(1 - \mathrm{SSIM}\left(\widehat{\mathbf{I}}_{\mathrm{f}},\ \mathbf{I}^{\mathrm{vi}}\right)\right), \tag{31}$$

where $\widehat{\mathbf{I}}_{\mathrm{f}}$ is the fused image and β_1 is a positive parameter. Moreover, the gradient loss function is given as follows:

$$\mathcal{L}_{\mathrm{g}} = \beta_2\|\mathrm{SF}\left(\widehat{\mathbf{I}}_{\mathrm{f}}\right) - \mathrm{SF}\left(\mathbf{I}^{\mathrm{vi}}\right)\|_2 + \beta_3\|\mathrm{SF}\left(\widehat{\mathbf{I}}_{\mathrm{f}}\right) - \mathrm{SF}\left(\mathbf{I}^{\mathrm{ir}}\right)\|_2, \tag{32}$$

where β_2, β_3 are two positive parameters, $\|\cdot\|_2$ stands for the ℓ_2 norm, and $\mathrm{SF}(\cdot)$ indicates the spatial frequency (SF) operation (Eskicioglu & Fisher, 1995), which mainly reflects texture details of the image, and is calculated as:

$$\mathrm{SF} = 1 - \sqrt{\mathrm{Hor}^2 + \mathrm{Ver}^2}, \tag{33}$$

where "Hor" and "Ver" denote the horizontal and vertical gradients, respectively. The overall forward loss function for the IVF task is as follows:

$$\mathcal{L}_{\mathrm{f}}^{IVF} = \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{g}}. \tag{34}$$

4 Experiments

This section is devoted to some experiments on three representative fusion tasks, i.e., pansharpening, HMF, and IVF[11] to validate the effectiveness of the given DCINN method comparing with some recent SOTA approaches. We will exhibit first the experiment settings (including datasets, metrics, and compared methods) for each task, then quantitatively and qualitatively comparing the given method with other competitive approaches on several examples. Moreover, we assess the effectiveness of our method with an ablation study, as well as including a discussion about the hyperparameter λ, which is a key parameter for our approach.

Note that our model is coded with Pytorch 1.12.1 and trained on an NVIDIA GeForce RTX 3090 GPU. Besides, we use the Adam optimizer (Kingma & Ba, 2014) for optimization with the default setting. Moreover, the batch size for the training is 32 for all the tasks.

[9] The symbol "|" indicates that CINN takes $\mathbf{F}_{\mathrm{a}}$ as conditional features.

[10] The reason why using different learning ways is given in Sect. 3.1 Even though there are different learning ways, the incorporation into a uniform framework is not affected.

[11] IVF is considered a typical multi-model image fusion task addressed in an unsupervised way.

4.1 Pansharpening Experiments

4.1.1 Experimental Setup

Regarding the datasets, we used two benchmark remote sensing datasets acquired by WorldView-2 (WV2) and WorldView-3 (WV3)[12] to conduct our experiments. Note that since the GT images in pansharpening are not available, Wald's protocol (Wald et al., 1997) is used to generate the training and testing datasets at reduced resolution. The simulation process of the training and testing data is reported in (Deng et al., 2021). It mainly consists of three steps: (1) we downsample the original PAN and the original MS image by a scaling factor of 4 using modulation transfer function (MTF)-based filters, then we view the downsampled PAN image as the PAN image for training and the downsampled MS image as the LRMS image for training; (2) the original MS image is considered as the GT (or labeled) image for training; (3) we upsample the LRMS image by using a polynomial kernel with 23 coefficients (Aiazzi et al., 2002) to obtain the upsampled version of the MS image.

Regarding the parameter setting, the spatial sizes of the simulated LRMS and PAN images (or patches) for training are 16 × 16 pixels and 64 × 64 pixels, respectively. We simulated 11,322 image pairs and then splitting them into 8806 samples for training and 2516 samples for validation. The training and validation dataset can be found in (Deng et al., 2022). The hyperparameter λ of the loss function is set to 1. Besides, our model is trained with 350 epochs, and the initial learning rate is 10^{-3} with decays at the 50-th, 100-th, 250-th, and 300-th epoch by a factor of 2.

The quality metrics used to assess the performance at reduced-resolution are three widely-used ones: Q8 (Garzelli & Nencini, 2009), relative dimensionless global error synthesis (ERGAS) (Wald, 2002), and spectral angle mapper (SAM) (Yuhas et al., 1992). Q8 is an overall quality index measuring both radiometric and spectral distortions, SAM mainly measures the spectral distortion between the estimated image and the GT, and ERGAS reflects a radiometric distortion. For examples at full-resolution (i.e., using data without any simulation step), the quality with no reference (QNR) index (Alparone et al., 2008) is used consisting of a spectral, D_λ, and a spatial, D_S, distortions (Alparone et al., 2008). The ideal values are 1 for QNR and Q8, and 0 for ERGAS, SAM, D_λ, and D_S.

For benchmarking approaches, recent SOTA traditional and DL-based methods are chosen to evaluate the performance. Specifically, traditional methods include both representative component substitution methods (GS (Craig & Bernard, 2000), PRACS (Choi et al., 2010), BDSD (Andrea et al., 2007)) and classical MRA methods (as SFIM (Liu, 2002) and GLP-Reg (Vivone et al., 2018)). We also compare our DCINN with an optimization-based method, *i.e.,* LRTCFPan (Wu et al., 2023). Additionally, DL-based methods include recent representative approaches as PanNet (Yang et al., 2017), DMDNet (Fu et al., 2020), FusionNet (Deng et al., 2021), GPPNN (Xu et al., 2021), TDNet (Zhang et al., 2022), LACNet (Jin et al., 2022b), and PanFormer (Zhou et al., 2022). The selected DL-based methods are chosen because they reported the SOTA performance with code available. For completeness, we also compare our method with two classical DL-based methods, i.e., PNN (Giuseppe et al., 2016) and DiCNN (He et al., 2019). Note that all the compared DL-based methods are retrained with default parameters on the same training dataset for a fair comparison.

4.1.2 Quantitative Results

We generated 78 samples for the reduced-resolution assessment and 200 samples for the full-resolution assessment. For both the datasets, the PAN size is 256 × 256. The results for both the assessment (at reduced and full resolution) are reported in Table 2. As we can see, our DCINN achieves the best performance considering both the reduced-resolution and full-resolution assessments. Specifically, DCINN obtains the first place considering all the three metrics on the reduced-resolution experiments. The DL-based methods generally perform better than the traditional methods. Besides, the DL-based methods achieve similar performance at full-resolution except for TDNet (Zhang et al., 2022). While our DCINN performs much better than the other DL-based methods at full-resolution thanks to the invertibility that can better preserve the spatial and spectral information. Besides, we also reported the parameter amount of DL methods in the last column of Table 2. As we can see, the parameter amount of DCINN is similar to GPPNN (Xu et al., 2021) and TDNet (Zhang et al., 2022), but larger than the other DL-based methods. Overall, the parameter amount of most of the selected DL-based methods is comparable.

4.1.3 Visual Results

The visual comparison results on reduced-resolution examples are shown in Figs. 7 and 8. From the error maps, it is clear that the proposed DCINN achieves minimal reconstruction loss. Moreover, we also report the visual results for the full-resolution dataset in Fig. 9. From the close-ups (green rectangles), all the compared DL-based methods (except for DMDNet) get striped distortion, indicating that the spatial fidelity of these methods is not good. Compared with the other approaches, our method gets the best visual performance, i.e., better color (spectral) preservation and clearer textures.

[12] http://www.digitalglobe.com/samples?search=Imagery.

Table 2 Average quantitative outcomes for all the compared approaches on 78 reduced-resolution examples and 200 full-resolution examples from the WV3 dataset

Method	Reduced-Resolution			Full-Resolution			#Param. (M)
	SAM↓	ERGAS↓	Q8↑	QNR↑	D_λ↓	D_s↓	
GS (Craig & Bernard, 2000)	5.526±2.598	6.223±2.692	0.675±0.223	0.902±0.045	0.017±0.019	0.082±0.032	\
SFIM (Liu, 2002)	4.934±2.283	6.053±3.540	0.712±0.229	0.934±0.038	0.021±0.021	0.045±0.021	\
BDSD (Andrea et al., 2007)	5.375±2.491	5.586±2.448	0.718±0.241	0.930±0.027	0.019±0.009	0.050±0.021	\
PRACS (Choi et al., 2010)	5.133±2.470	5.791±2.537	0.698±0.231	0.914±0.044	0.017±0.016	0.069±0.032	\
GLP-Reg (Vivone et al., 2018)	4.896±2.405	5.193±2.298	0.734±0.236	0.919±0.049	0.021±0.023	0.054±0.031	\
PNN (Giuseppe et al., 2016)	3.521±1.252	3.066±1.222	0.803±0.259	***0.959±0.026***	0.016±0.014	0.025±0.013	0.31M
DiCNN1 (He et al., 2019)	3.411±1.283	2.998±1.062	0.810±0.254	0.946±0.032	0.016±0.016	0.038±0.020	***0.18M***
PanNet (Yang et al., 2017)	3.231±1.210	2.899±1.052	0.811±0.254	0.958±0.019	0.022±0.010	**0.020±0.011**	0.25M
DMDNet (Fu et al., 2020)	3.070±1.117	***2.716±0.984***	0.815±0.254	0.955±0.021	***0.014±0.012***	0.030±0.012	0.32M
FusionNet (Deng et al., 2021)	***3.053±1.104***	2.750±0.973	***0.818±0.250***	0.956±0.027	0.017±0.015	***0.020±0.016***	0.23M
GPPNN (Xu et al., 2021)	3.059±1.039	2.755±0.975	0.811±0.261	0.951±0.024	0.016±0.015	0.032±0.012	0.238M
TDNet (Zhang et al., 2022)	3.351±1.071	2.919±1.069	0.803±0.287	0.925±0.041	0.031±0.025	0.045±0.020	0.55M
LACNet (Jin et al., 2022b)	3.132±1.171	2.853±1.163	0.811±0.260	0.938±0.030	0.023±0.019	0.039±0.015	**0.054M**
PanFormer (Zhou et al., 2022)	3.155±1.045	4.650±4.451	0.813±0.258	0.953±0.027	0.016±0.014	0.030±0.015	0.279M
LRTCFPan (Wu et al., 2023)	4.443±1.996	4.815±2.182	0.754±2.182	0.944±0.039	0.020±0.017	0.036±0.025	\
DCINN	**2.829±1.025**	**2.447±0.862**	**0.821±0.257**	**0.965±0.016**	**0.012±0.008**	0.023±0.009	0.501M

The best and second results are in bold and bolditalic, respectively

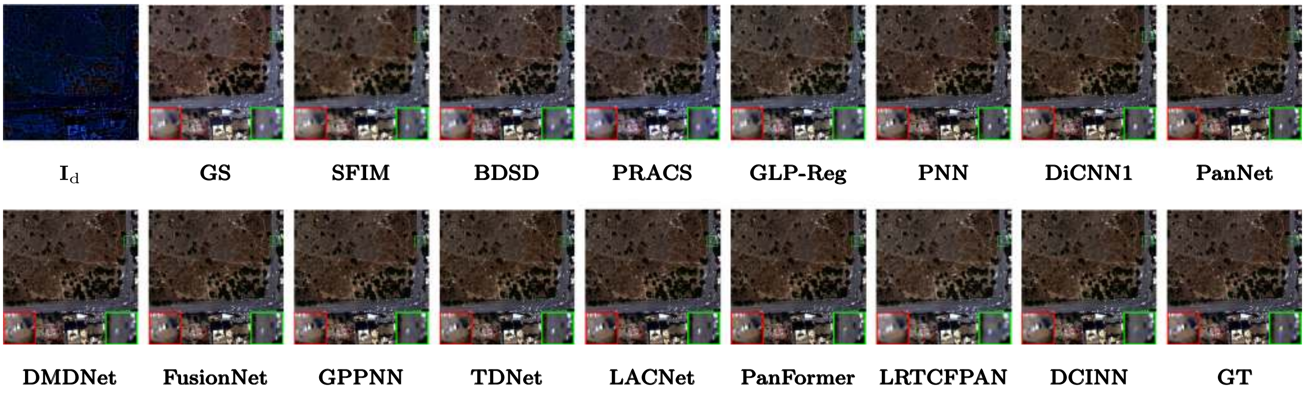


Fig. 7 Visual comparisons of all the compared approaches on the reduced-resolution WV3 dataset

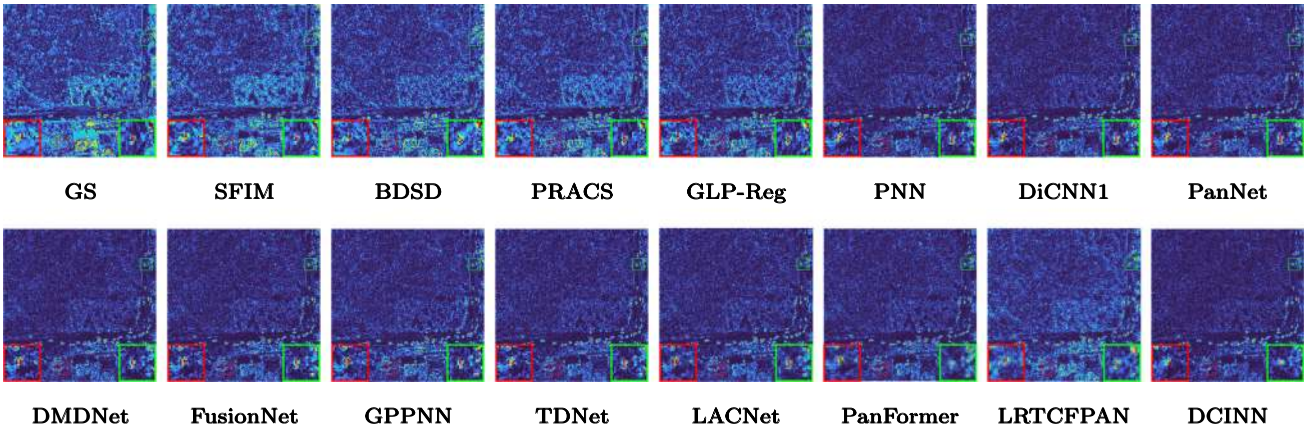


Fig. 8 The corresponding absolute error maps (AEMs) on the reduced-resolution WV3 dataset. For simplicity, only band 5 is shown

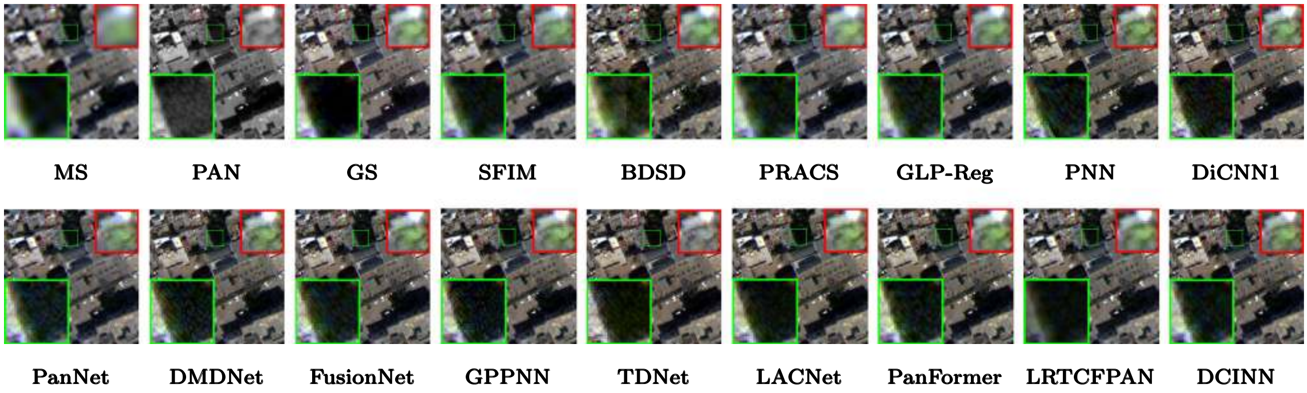


Fig. 9 Visual comparisons of all the compared approaches on the full-resolution WV3 dataset

Table 3 Average quantitative outcomes of all the compared approaches on 20 samples from WV2 dataset

Method	SAM↓	ERGAS↓	Q8↑
GS (Craig & Bernard, 2000)	6.6657	5.2616	0.7734
SFIM (Liu, 2002)	6.1848	4.7331	0.8103
BDSD (Andrea et al., 2007)	9.5727	6.4377	0.7522
PRACS (Choi et al., 2010)	6.3483	5.3019	0.7668
GLP-Reg (Vivone et al., 2018)	6.2257	4.4341	0.8278
PNN (Giuseppe et al., 2016)	7.0493	5.0363	0.8179
DiCNN1 (He et al., 2019)	6.1902	5.2940	0.7889
PanNet (Yang et al., 2017)	***5.3886***	4.2670	0.8398
DMDNet (Fu et al., 2020)	**5.2642**	4.1260	0.8493
FusionNet (Deng et al., 2021)	6.1348	5.0248	0.7922
GPPNN (Xu et al., 2021)	6.8776	4.8423	0.8246
TDNet (Zhang et al., 2022)	9.2302	6.0426	0.8082
LACNet (Jin et al., 2022b)	7.0453	5.0363	0.8179
PanFormer (Zhou et al., 2022)	5.4974	4.5421	0.8304
LRTCFPan (Wu et al., 2023)	5.6297	***4.1095***	***0.8502***
DCINN	5.4583	**3.9936**	**0.8639**

The best and second results are in bold and bolditalic, respectively

4.1.4 Network Generalization

To evaluate the generalization ability, the DL-based models trained on the WV3 dataset are further tested on the WV2 dataset, whose results are shown in Table 3. For a fair comparison, we directly adopt the WV2 dataset from the PanCollection (Deng et al., 2022), which includes 20 challenging samples and is available.[13] Among the compared methods, the LDM-based techniques, i.e., PanNet (Yang et al., 2017), DMDNet (Fu et al., 2020), and FusionNet (Deng et al., 2021), demonstrate a better generalization ability than other methods. The good generalization ability of LDM-based methods comes from the fact that the complexity of the learned mapping (only for image details) is smaller than that of LIM-based methods, which cannot lead to an important overfitting. From the table, our DCINN achieves comparable performance with the DMDNet, which showed SOTA network generalization ability in the related literature.

4.2 HMF Experiments

4.2.1 Experimental Setup

This section mainly performs HMF experiments on CAVE [14] and Harvard (Chakrabarti & Zickler, 2011) datasets. The CAVE dataset contains 32 HS images with 31 spectral bands, and each band has a spatial size of 512 × 512. In the experiment, we select 21 HS images for training and the rest 11 HS images are used for testing. Besides, the Harvard dataset contains 77 HS images with 31 spectral bands, and each band has a spatial size of 1392 × 1040. For this dataset, we randomly select 10 HS images for testing and the rest 67 HS images are used for training. Since both the datasets do not contain LRHS and HRMS images, the original HS images are taken as reference images (i.e., HRHS images) and LRHS and HRMS images are generated through simulation. Before the simulation phase, the original HS images are normalized to [0, 1] by dividing them by $2^{16} - 1$ because we work with 16-bits data.

For data simulation, we use first the spectral response matrix of the Nikon D700[15] camera to generate HRMS images. Following previous methods as (Dian et al., 2018), the original spectral response matrix is normalized such that each column sums to one. The LRHS images are generated by blurring the reference image by a zero-mean 8 × 8 Gaussian blur kernel with variance equal to 1, and, afterwards, downsampling with the "nearest" interpolation.

[13] https://github.com/liangjiandeng/PanCollection.

[14] https://www1.cs.columbia.edu/CAVE/databases/multispectral/.

[15] https://maxmax.com/nikon_d700_study.

For the parameter setting, the spatial sizes of the simulated LRHS and HRMS image patches for training are 10 × 10 pixels and 40 × 40 pixels for a scaling factor of 4, and 10 × 10 pixels and 80 × 80 pixels for a scaling factor of 8. The model is trained for 350 epochs on both CAVE and Harvard datasets. The initial learning rate is 10^{-3} and is decayed in the 50-th, 100-th, 250-th, and 300-th epoch by a factor of 2.

For quality metrics, we considered three widely-used ones, i.e., peak signal-to-noise ratio (PSNR), ERGAS (Wald, 2002), and SAM (Yuhas et al., 1992). We also evaluate the parameter amount for each compared DL-based method. The ideal value for the PSNR index is $+\infty$.

Regarding the compared approaches, some recent SOTA traditional and DL-based methods are chosen. The traditional methods include Fuse (Qi et al., 2015), CNMF (Naoto et al., 2012), Lanaras (Lanaras et al., 2015), and HySure (Miguel et al., 2015). Instead, the considered DL-based methods are ResTFNet (Liu et al., 2020b), SSRNet (Zhang et al., 2021), HSRNet (Hu et al., 2022b), DBIN (Wang et al., 2019), Mog-DCN (Dong et al., 2021), DHIF (Huang et al., 2022), Fusformer (Hu et al., 2022a), and 3DTNet (Ma et al., 2023). We retrain all the compared DL-based methods on the same training dataset with their default setting for fair comparison.

4.2.2 Quantitative Results

We report first the experimental results for the scaling factor of 8 on the CAVE dataset. As we can see from Table 4, the performance of our DCINN are better than the other methods considering all the quality metrics. Specifically, compared with 3DTNet (Ma et al., 2023), the PSNR improvement is about 0.83 dB. Since DL-based methods perform much better than traditional methods, we only show the results of DL-based methods in the following experiments on the HMF task. The quantitative experimental results for a scaling factor of ×4 on Harvard and CAVE are reported in Table 5. Again,

Table 4 Average metrics for all the compared approaches on 11 samples of the CAVE dataset for a scaling factor of 8

Method	PSNR↑	SAM↓	ERGAS↓
Fuse (Qi et al., 2015)	36.92±4.468	8.77±4.83	2.64±1.783
CNMF (Naoto et al., 2012)	36.54±5.56	7.19±3.23	3.33±3.674
Lanaras (Lanaras et al., 2015)	36.93±4.197	6.85±2.646	2.56±1.645
HySure (Miguel et al., 2015)	36.86±4.668	11.66±5.593	2.83±2.299
ResTFNet (Liu et al., 2020b)	43.74±5.349	3.53±0.925	2.75±2.521
SSRNet (Zhang et al., 2021)	46.21±4.194	3.13±0.9704	2.08±1.454
HSRnet (Hu et al., 2022b)	46.68±4.476	2.91±0.856	1.85±1.267
Fusformer (Hu et al., 2022a)	47.95±7.794	2.74±1.295	1.42±2.619
DBIN (Wang et al., 2019)	48.96±4.742	2.53±0.726	0.78±0.658
Mog-DCN (Dong et al., 2021)	49.17±5.000	2.49±0.734	0.75±0.637
DHIF (Huang et al., 2022)	48.46±4.893	2.50±0.787	0.84±0.672
3DTNet (Ma et al., 2023)	***49.22±4.770***	***2.39±.683***	***0.73±0.652***
DCINN	**50.05±5.324**	**2.28±0.711**	**0.71±0.680**

The best and second results are in bold and bolditalic, respectively

Table 5 Average metrics for all the DL-based approaches on 11 samples of the CAVE dataset and 10 samples of the Harvard dataset for a scaling factor of ×4

Method	CAVE			Harvard			#Param. (M)
	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓	
ResTFNet (Liu et al., 2020b)	45.58±5.465	2.82±0.700	2.36±2.587	45.93±4.352	2.61±0.693	2.56±1.319	2.387M
SSRNet (Zhang et al., 2021)	48.62±3.918	2.54±0.837	1.63±1.206	47.95±3.368	2.31±0.604	2.30±1.417	**0.027M**
HSRnet (Hu et al., 2022b)	50.38±3.380	2.23±0.658	1.20±0.750	***48.29±3.030***	2.26±0.557	***1.87±0.809***	0.587M
Fusformer (Hu et al., 2022a)	49.98±8.097	2.20±0.851	1.25±2.603	47.87±5.125	2.84±2.069	2.04±0.988	0.504M
DBIN (Wang et al., 2019)	50.83±4.293	2.21±0.627	1.24±1.059	47.88±3.869	2.31±0.460	1.95±0.813	***0.469M***
Mog-DCN (Dong et al., 2021)	***51.63±4.097***	***2.026±0.615***	***1.11±0.820***	47.89±4.097	***2.11±0.523***	1.89±0.823	7.071M
DHIF (Huang et al., 2022)	51.07±4.165	2.01±0.630	1.22±0.967	47.68±3.849	2.32±0.528	1.95±0.915	22.609M
3DTNet (Ma et al., 2023)	51.38±4.179	2.16±0.695	1.14±0.996	47.78±4.423	2.04±0.509	1.98±0.858	3.46M
DCINN	**52.21±4.246**	**1.928±0.614**	**1.043±0.843**	**49.35±3.276**	**2.04±0.518**	**1.74±0.815**	4.32M

The best and second results are in bold and bolditalic, respectively

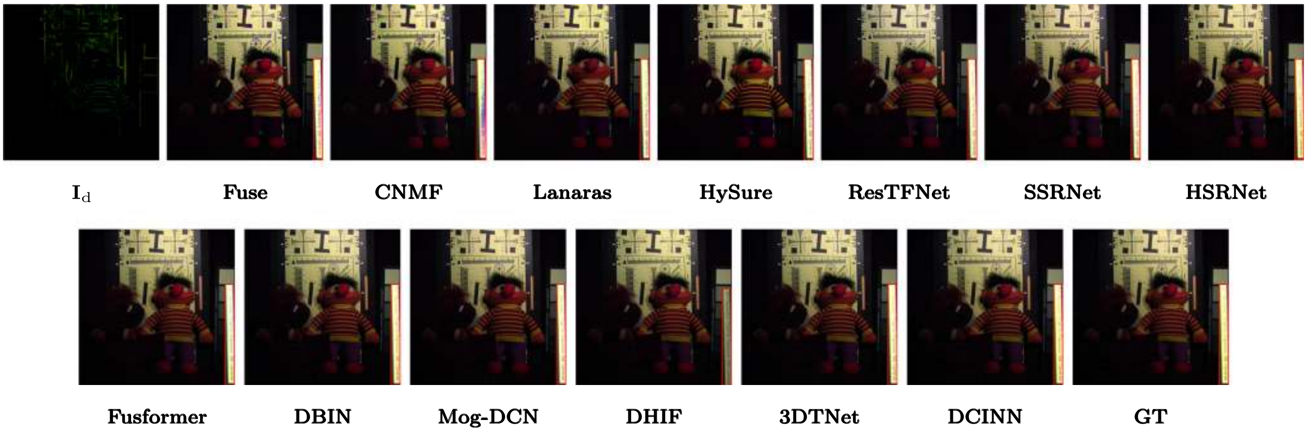


Fig. 10 Visual comparisons of all the compared approaches on the CAVE dataset. The false color images are generated by taking the 30th, 15th, and 2nd bands of the fused outcomes as red (R), green (G), and blue (B) channels, respectively

DCINN consistently outperforms the other methods. More specifically, the PSNR gains on CAVE and Harvard data are 0.585 dB and 1.06 dB, respectively. Moreover, the parameter amount of DCINN is less than Mog-DCN (Dong et al., 2021) and much less than DHIF (Huang et al., 2022) (see the last column of Table 5), which verifies that DCINN is not only effective but also efficient. These experiments demonstate that DCINN can better preserve the spectral and spatial information than the compared methods. It is worth noting that the Harvard dataset is considered to be much harder than CAVE dataset due to the heavy noise on the LRHS images. The performance gain on the Harvard dataset clearly verifies that DCINN is also robust to the noise.

4.2.3 Visual Results

The visual experimental results with the scaling factor of 8 are shown in Fig. 10. As can be seen from the rectangular box, our method yields the most accurate color, related to the ground-truth image. Since it is hard to distinguish the differences in image visual quality among the DL-based methods, we also showed the corresponding absolute error maps (AEMs) in Fig. 11. According to the error maps, our method obtains the smallest reconstruction error.

4.2.4 Network Generalization

To evaluate the generalization ability, the DL-based models trained on the Harvard dataset with a scaling factor of ×4 are further tested on the CAVE dataset with the same scaling factor, whose results are shown in Table 6. From the table, our DCINN obtains the second-best performance on PSNR and ERGAS metrics while HSRNet (Hu et al., 2022b) achieves the best performance.

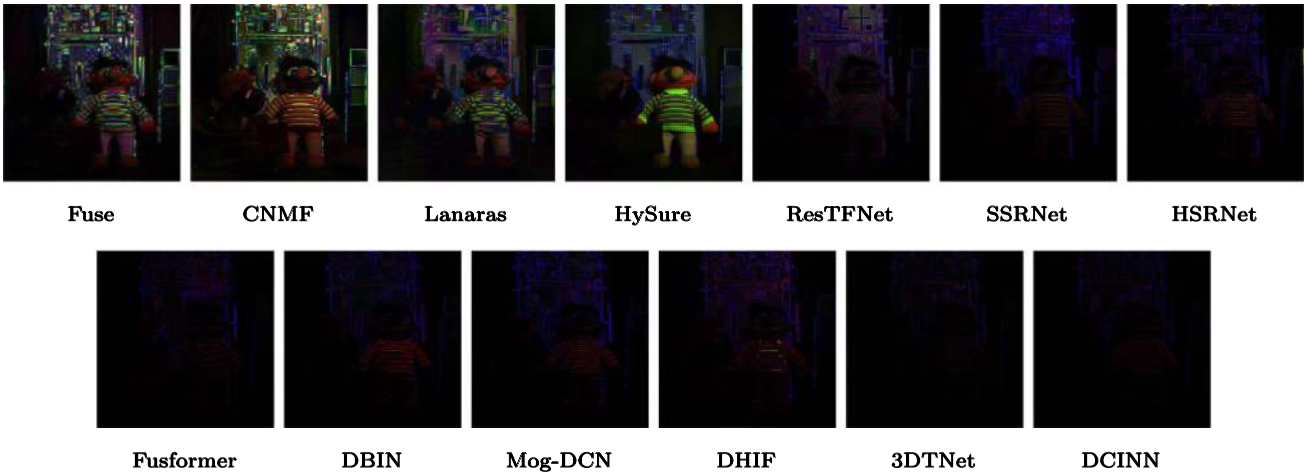


Fig. 11 The corresponding absolute error maps using the reference (GT) image on the CAVE dataset. For clarity, the image intensity of the original error maps are magnified by a factor of 5

Table 6 Average quantitative outcomes of all the compared approaches trained on the Harvard dataset and tested on 11 samples of the CAVE dataset

Method	PSNR↑	SAM↓	ERGAS↓
ResTFNet (Liu et al., 2020b)	25.14	21.78	20.67
SSRNet (Zhang et al., 2021)	36.88	21.98	9.88
HSRNet (Hu et al., 2022b)	**42.85**	**5.33**	**2.56**
FusFormer (Hu et al., 2022a)	35.44	***6.11***	6.75
DBIN (Wang et al., 2019)	35.55	24.49	32.88
Mog-DCN (Dong et al., 2021)	33.47	10.83	8.80
DHIF (Huang et al., 2022)	25.28	24.09	22.83
3DTNet (Ma et al., 2023)	24.18	10.37	20.65
DCINN	***36.75***	8.06	***5.25***

The best and second results are in bold and bolditalic, respectively

4.3 IVF Experiments

4.3.1 Experimental Setup

Regarding the IVF experiments, we used two benchmark datasets, i.e., RoadScene (RS) (Xu et al., 2020) and TNO (Alexander, 2017). The RS dataset is used for both training and testing, and the TNO dataset is only used for testing. Specifically, the RS dataset contains 221 pairs of aligned infrared and RGB-visible images. We choose 201 samples for training and 20 samples for testing. Besides, we selected 8 samples of TNO for testing. Following previous methods such as in (Tang et al., 2022), the visible images are converted first into the YCbCr color space and only the Y (luminance) channel is used for fusion.

About the parameter setting, we random crop image patches with spatial size of 128 × 128 for training. Because the visible (VI) images have more details than the infrared (IR) images on the TNO dataset while in the RS dataset we have the opposite behavior, we train two models for each dataset. The hyperparameters of the loss functions are set as $\beta_1 = 0.05$, $\beta_2 = 6 \times 10^{-3}$, and $\beta_3 = 2.5 \times 10^{-3}$ for the RS dataset and $\beta_1 = 1$, $\beta_2 = 1 \times 10^{-3}$, and $\beta_3 = 0$ for the TNO dataset, respectively. We use the mean rule and the max rule for the fusion of base components in the case of the RS and TNO datasets, respectively. Our model is trained with 20 epochs with a learning rate of 5×10^{-5}.

Regarding the quality metrics, we selected four commonly-used indexes, i.e., entropy (EN) (Wesley et al., 2008), mutual information (MI) (Qu et al., 2002), standard deviation (SD) (Rao, 1997), and the multi-scale structural similarity index (MS-SSIM) (Wang et al., 2003). The ideal values are $+\infty$ for EN, MI, and SD, and 1 for MS-SSIM. Both EN (Wesley et al., 2008) and SD (Rao, 1997) are no-reference metrics. Specifically, EN measures the amount of information in the fused image, while SD reflects the distribution and contrast of an image. In addition, MS-SSIM (Wang et al., 2003) and MI (Qu et al., 2002) are used as semi-reference metrics[16] calculated by comparing the fused outcome with the source images. MS-SSIM (Wang et al., 2003) calculates the structural similarity between the fused image and the source images, while MI (Qu et al., 2002) computes the mutual information between the fused image and the source images.

Finally, about the compared approaches, nine traditional and DL-based methods reporting SOTA performance are selected for evaluation. Traditional methods include NSCT (Adu et al., 2013) and GTF (Ma et al., 2016). Moreover, some representative DL-based methods as LRRNet (Li et al., 2023), SwinFusion (Ma et al., 2022), YDTR (Tang et al., 2022), RFN-Nest (Li et al., 2021), DenseFuse (Li & Wu, 2019), IFCNN (Zhang et al., 2020), and U2Fusion (Xu et al., 2022) are considered.

4.3.2 Visual Results

The qualitative results on the RS dataset are shown in Fig. 12. From the rectangular boxes in Fig. 12, GTF (Ma et al., 2016), RFN-Nest (Li et al., 2021), and SwinFusion (Ma et al., 2022) show outcomes with blurring effects (see the red rectangular box), the fused outcomes of DenseFuse (Li & Wu, 2019) and U2Fusion (Xu et al., 2022) cannot retain background details (see the blue rectangular boxes), YDTR (Tang et al., 2022) generates artifacts (see the blue rectangular box). LRRNet (Li et al., 2023) generates blur salient targets (see the green rectangular box). Instead, DCINN not only preserves both details from VI images and salient targets from IR images, but also produces less distortion. The visual results on the TNO dataset are also depicted in Fig. 13. From Fig. 13, we can draw a similar conclusion as for the RS dataset. DCINN generally yields very competitive visual quality.

4.3.3 Quantitative Results

The quantitative results are reported in Table 7. For the RS dataset, DCINN obtains the best performance considering all the metrics. GTF (Ma et al., 2016) performs very competitively on EN (Wesley et al., 2008), MI (Qu et al., 2002), and SD (Rao, 1997), however, it achieves the worst MS-SSIM (Wang et al., 2003). For the TNO dataset, DCINN ranks second for the MS-SSIM (Wang et al., 2003) and SD (Rao, 1997) metrics. Although YDTR (Tang et al., 2022) obtains satisfactory quantitative results, its fused outcomes contain significant artifacts (see Sect. 4.3.2). The promising quantitative results verify that DCINN can better persevere details. Besides, the parameter amount of DCINN is smaller than U2Fusion (Xu et al., 2022) but larger than the other DL-

[16] Both MS-SSIM and MI can be used as reference metrics, but for the IVF task, the references are not available.

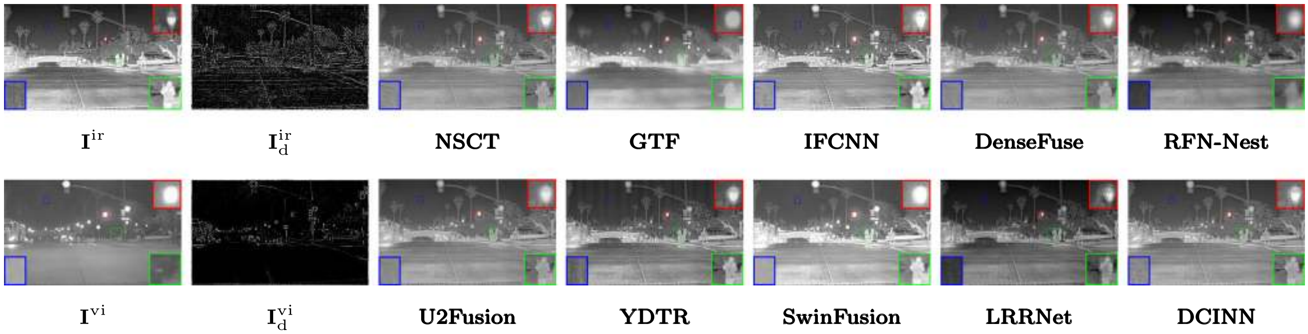

$\mathbf{I}^{\mathrm{ir}}$
$\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}}$
NSCT
GTF
IFCNN
DenseFuse
RFN-Nest
$\mathbf{I}^{\mathrm{vi}}$
$\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}$
U2Fusion
YDTR
SwinFusion
LRRNet
DCINN

Fig. 12 Visual comparisons of all the compared approaches on the RoadScene dataset

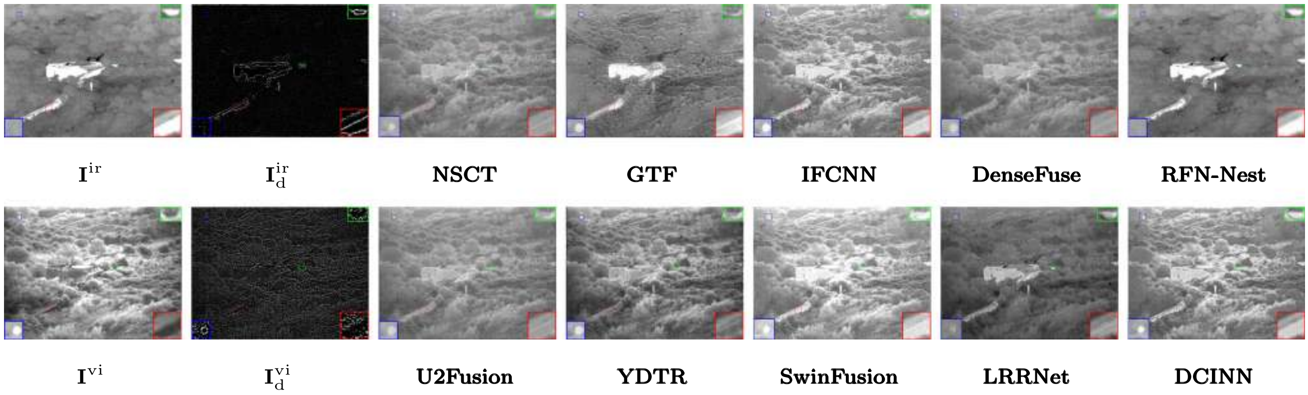

$\mathbf{I}^{\mathrm{ir}}$
$\mathbf{I}_{\mathrm{d}}^{\mathrm{ir}}$
NSCT
GTF
IFCNN
DenseFuse
RFN-Nest
$\mathbf{I}^{\mathrm{vi}}$
$\mathbf{I}_{\mathrm{d}}^{\mathrm{vi}}$
U2Fusion
YDTR
SwinFusion
LRRNet
DCINN

Fig. 13 Visual comparisons of all the compared approaches on the TNO dataset

Table 7 Average metrics of all the compared DL-based approaches on 20 samples of the RoadScene dataset and 8 samples of the TNO dataset, respectively

Method	RoadScene				TNO				#Param. (M)
	EN↑	MI↑	SD↑	MS-SSIM↑	EN↑	MI↑	SD↑	MS-SSIM↑	
NSCT (Adu et al., 2013)	6.8707	13.7414	54.4155	0.8424	6.2322	12.4644	45.7987	0.8856	\
GTF (Ma et al., 2016)	7.2658	14.5316	***74.6608***	0.7009	6.3643	12.7287	52.8844	0.8264	\
IFCNN (Zhang et al., 2020)	7.0541	14.1082	62.3081	0.8418	6.5908	13.1816	63.3083	**0.9171**	***0.083M***
RFN-Nest (Li et al., 2021)	7.1470	14.2939	67.5769	0.7859	6.6243	13.2486	60.2526	0.6037	30.096M
DenseFuse (Li & Wu, 2019)	6.8248	13.6496	53.2366	0.8289	6.1828	12.3656	44.4600	0.8777	0.890M
U2Fusion (Xu et al., 2022)	6.8554	13.7107	54.2737	0.8349	6.2173	12.4347	45.5796	0.8264	2.636M
YDTR (Tang et al., 2022)	7.2438	14.4877	70.9238	0.8149	**6.7668**	**13.5335**	67.2590	0.8407	0.871M
SwinFusion (Ma et al., 2022)	7.1779	14.3559	69.4970	***0.8490***	6.5979	13.1958	**72.6153**	***0.9061***	0.973M
LRRNet (Li et al., 2023)	***7.2684***	***14.5369***	60.9503	0.7803	6.6269	13.2539	51.4271	0.7263	**0.049M**
DCINN	**7.3011**	**14.6021**	**75.9048**	**0.8653**	***6.6371***	***13.2743***	***67.9157***	0.8896	2.521M

The proposed DCINN uses the mean rule and max rule for the fusion of base components on the RS and TNO datasets, respectively. The best and second results are in bold and bolditalic, respectively

based methods. The parameter amount of DCINN could be potentially reduced by using fewer Inv Blocks.

4.4 Ablation Study and Discussion

This section is devoted to a series of ablation studies to verify the effectiveness of each contribution of the proposed DCINN method. Moreover, we will discuss the influence of a key hyperparameter in our given method. Besides, we will study the effects of the fusion rules on the performances for the IVF task.

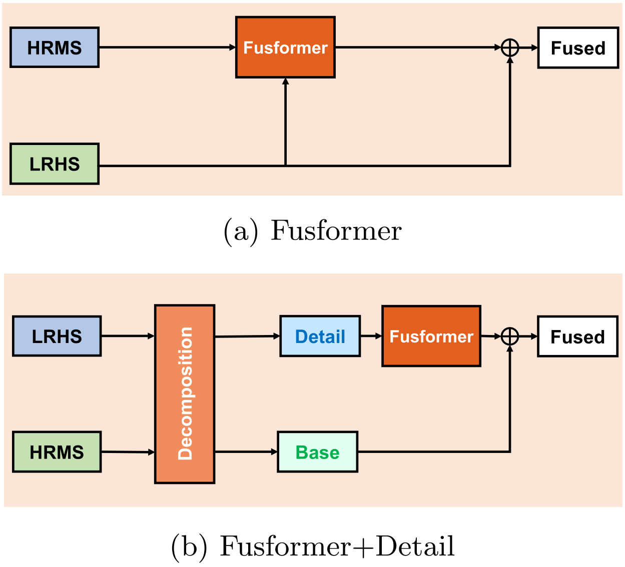
4.4.1 Effect of the ANet

Unlike previous LDM-based methods that only use details for inference, we design an auxiliary network to make full

Table 8 Ablation experiments on the three image fusion tasks

	Pansharpening (WV3 ×4)			HMF (CAVE ×8)			IVF (RS)			
	SAM↓	ERGAS↓	Q8↑	PSNR↑	SAM↓	ERGAS↓	EN↑	MI↑	SD↑	MS-SSIM↑
W/O ANet	***2.8861±1.042***	2.536±0.906	0.82±0.255	***49.56±5.603***	***2.39±0.751***	***0.77±0.846***	6.9430	13.8861	56.4254	0.8591
W/O Detail	2.897±1.040	***2.534±0.928***	**0.822±0.254**	48.68±4.959	2.67±0.792	0.82±0.705	***7.2896***	***14.5793***	***75.3216***	0.8457
W/O CINN	2.9829±1.040	2.740±1.166	0.819±0.256	48.52±5.193	2.65±0.815	0.82±0.781	6.9173	13.8346	56.0078	**0.8747**
DCINN	**2.829±1.025**	**2.447±0.862**	***0.821±0.257***	**50.05±5.324**	**2.28±0.711**	**0.71±0.680**	**7.3011**	**14.6021**	**75.9048**	***0.8653***

The best and second results are in bold and bolditalic, respectively



(a) Fusformer

(b) Fusformer+Detail

Fig. 14 **a** The original structure of Fusformer; **b** Our version of Fusformer

use of the information of the source images to help the inference. To assess its validity, we remove the ANet and report the experimental results, which are marked as “W/O ANet". Table 8 shows that its performance is obviously worse than the one of the proposed DCINN considering all the metrics on the three image fusion tasks. As shown by the first row of Fig. 15, without the ANet, the fused image shows a clear spatial distortion. This result indicates that using only details for reasoning is not enough and the ANet proposed in our method can help to exploit more information for reasoning, thus getting better performance.

4.4.2 Effect of CINN

To verify the effectiveness of our CINN, we conducted an experiment where we replaced the CINN with ResNet (He et al., 2016) to learn detail mapping. Meanwhile, we also eliminate the auxiliary network. For a fair comparison, we set the parameters of ResNet (He et al., 2016) to be roughly the same as our CINN. This model is marked with “W/O CINN”. Table 8 reports the experimental results showing that replacing CINN with ResNet (He et al., 2016) leads to a huge performance drop. As shown by the third row of Fig. 15, the fused image shows a large reconstruction error on the edges, which points out that CINN can help to preserve details with its lossless ability.

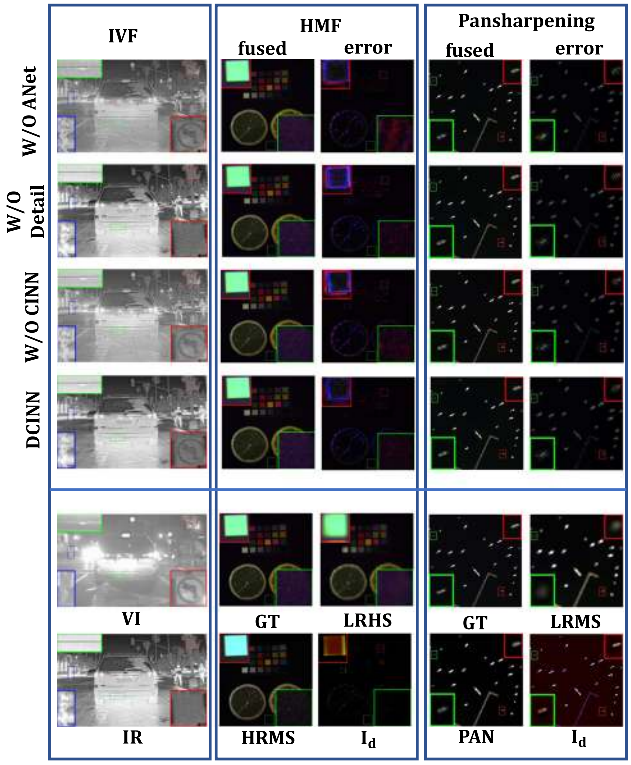


Fig. 15 Comparison of the fused images under the different settings

4.4.3 Effect of LDM

Unlike most of the previously developed methods, we propose to learn the detail mapping rather than LIM. To validate its effectiveness, we delete the ANet applying INN with LIM. As mentioned before, the INN requires the volume-preserving, however, this requirement cannot be satisfied for

Table 9 Average metrics for the Fusformer (Hu et al., 2022a) on 11 samples of the CAVE dataset for a scaling factor of ×4

Method	PSNR↑	SAM↓	ERGAS↓
Fusformer	***49.98±8.097***	**2.20±0.851**	***1.25±2.603***
Fusformer+Detail	**50.70±3.643**	***2.22±0.666***	**1.20±0.775**
DCINN	52.21±4.246	1.928±0.614	1.04±0.843

The best and second results are in bold and bolditalic, respectively

the involved image fusion tasks. To solve this problem, we propose two strategies for two different cases.

*Case*1 : For the IVF task, where the total size of the source images is bigger than the fused image, we adopt the strategy that has been widely used for the image denoising (Huang & Dragotti, 2022), the image rescaling (Xiao et al., 2020), and the image compression (Xu & Zhang, 2021). Specifically, the INN generates a fused image and a redundant variable that is assumed to follow a Gaussian distribution. The redundant variable serves to get the total size of outputs equal to the total size of inputs. We enforce the redundant variable following a Gaussian distribution to make sure that it contains no information related to the source images.

*Case*2 : For the pansharpening and HMF tasks, where the size of the source images is much smaller than the target image, there is no work to take a cue. In this paper, we propose to resize and replicate the source images to make their total sizes the same as the size of the target image.

This variant model is marked as "W/O Detail". According to Table 8 and Fig. 15, this variant achieves much worse performance than DCINN in both the cases. For the first case, as shown by the second row of the first column in Fig. 15, the fused outcome losses important visible details, which is due to the fact that the redundant variable still carries important information. For the second case, the degenerated performance is due to the low capacity of INN, which is not enough to model the complex image mapping.

4.4.4 Effect of the Decomposition Module for HMF Task

Based on the observation model of the HMF task, we proposed a new method to transform the image mapping into a detail mapping as shown in Eqs. (7) and (14). Our method can use any DL-based architecture to learn the detail mapping. To study its novelty, we apply it to Fusformer (Hu et al., 2022a), which is a recent LIM-based method (see the flowchart in Fig. 14a). More specifically, we inserted Fusformer into our detail framework as shown in Fig. 14b, and Table 9 reports the quantitative results.[17] As we can see, simply using Fusformer to learn the proposed detail mapping achieves a PSNR gain of about 0.7dB. In addition, the standard deviation is decreased, which shows that learning the proposed detail mapping helps achieving more robust performance.

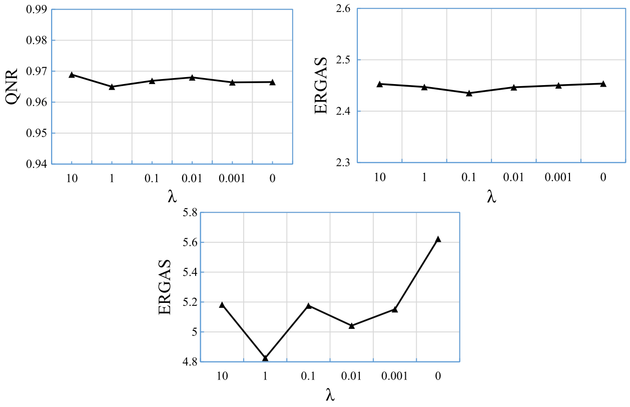


Fig. 16 The effect of λ on the full-resolution (using the QNR metric), the reduced-resolution (using the ERGAS metric), and the generalization performance (using the ERGAS metric) for the pansharpening task exploiting the WV3 dataset

4.4.5 Discussion on the Hyperparameter *λ*

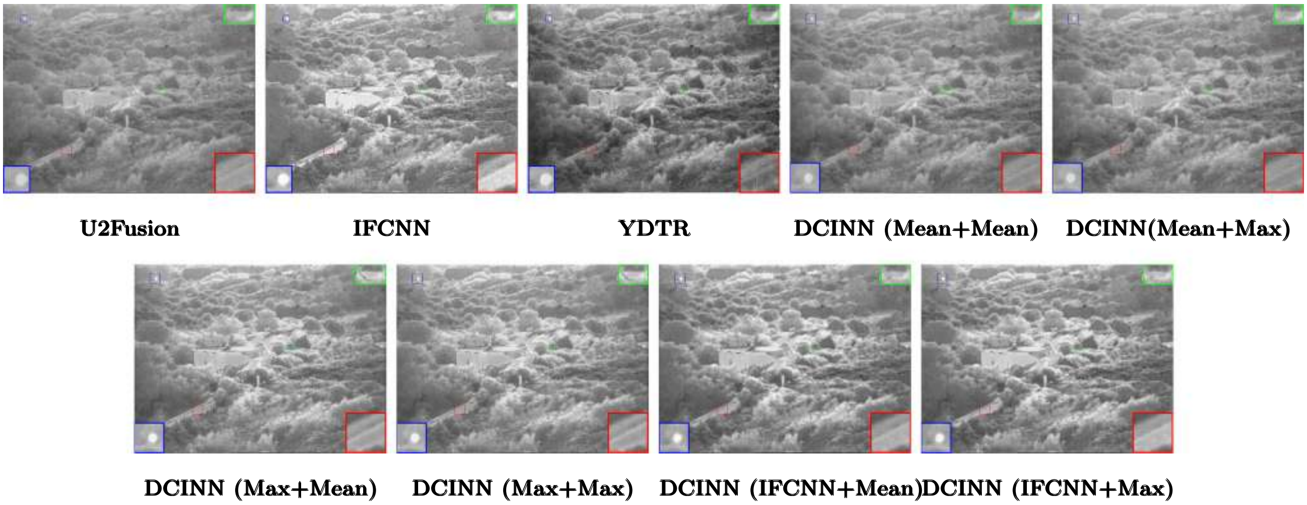
Recalling the loss function in Sect. 3.4, apart from the forward loss function to train DCINN on the pansharpening task, we also incorporate a backward loss function that enforces the reconstructed detail component to be consistent with the original detail component. The weight of the forward and backward losses are balanced by λ. To discuss the effectiveness of the key weighting parameter λ, we conducted several experiments at reduced-resolution and full-resolution, as well as showing the generalization performance, see Fig. 16. For simplicity, we only show the ERGAS at reduced-resolution and the QNR at full-resolution. λ has little influence at reduced-resolution and full-resolution but, instead, it has a great influence for the generalization ability. When $\lambda = 0$, DCINN achieves the worst performance, which implies that the backward loss improves the generalization performance, while DCINN achieves the best performance when $\lambda = 1$.

[17] DCINN can be just seen as an upper bound.

Table 10 Ablation experiments for the effects of fusion rules of the proposed DCINN on the TNO dataset

Base Rule	Detail Rule	EN↑	MI↑	SD↑	MS-SSIM↑
Mean	Mean	6.3465	12.6930	51.9685	0.9004
Mean	Max	6.3507	12.7014	52.0458	0.8991
Max	Mean	**6.6371**	**13.2743**	**67.9157**	0.8896
Max	Max	6.5333	13.0666	63.1392	0.9023
IFCNN	Mean	***6.6028***	***13.2056***	***67.1648***	**0.9109**
IFCNN	Max	6.5859	13.1718	67.0246	***0.9042***

The best and second results are in bold and bolditalic, respectively


U2Fusion IFCNN YDTR DCINN (Mean+Mean) DCINN(Mean+Max)

DCINN (Max+Mean) DCINN (Max+Max) DCINN (IFCNN+Mean)DCINN (IFCNN+Max)

Fig. 17 Visual comparisons of our DCINN using several fusion rules on the TNO dataset, where the caption “DCINN (Mean+Max)” denotes the DCINN with mean rule for the base fusion and max rule for detail fusion, respectively

Table 11 Average metrics of all the compared approaches on 20 samples of the Harvard medical image dataset for the task of MRI-CT image fusion

Method	EN↑	MI↑	SD↑	MS-SSIM↑
U2Fusion (Xu et al., 2022)	4.4808	8.9616	58.4926	0.8625
IFCNN (Zhang et al., 2020)	***4.4958***	***8.9916***	78.5851	0.9410
SwinFusion (Ma et al., 2022)	4.0171	8.0342	**89.7986**	0.9352
DCINN (Max+Mean)	4.3135	8.6270	***89.6357***	***0.9431***
DCINN (IFCNN+Mean)	**4.5304**	**9.0608**	82.7192	**0.9435**

The best and second results are in bold and bolditalic, respectively

4.4.6 Effect of the Fusion Rules for IVF Task

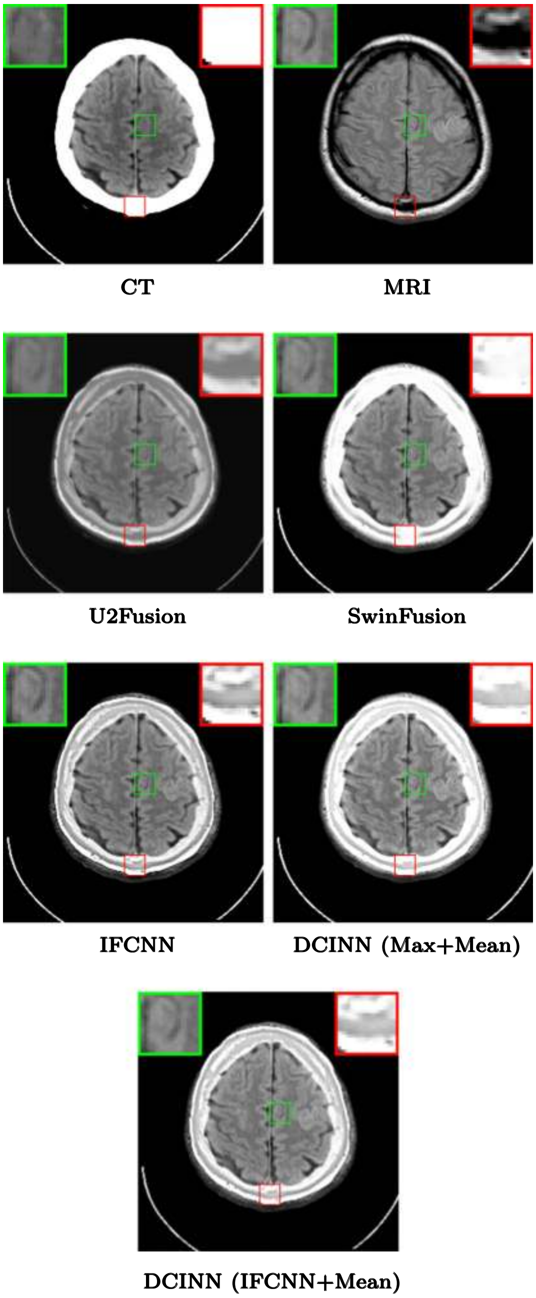
For the IVF problem, the final detail component and the base component are obtained by fusing the detail components and base components of the source images exploiting some fusion rules. To study the effects of the fusion rules, we conducted experiments on the TNO dataset. Table 10 reports the quantitative results. As we can see, using the max rule for base fusion significantly improves the performance. Moreover, using the SOTA fusion model, i.e., IFCNN (Zhang et al., 2020)[18] to fuse the base components also improves the performance. While using the mean rule or the max rule for the detail fusion has minimal effects on the performance. We also show the related results in Fig. 17. We can note that the fused products using the max rule or IFCNN (Zhang et al., 2020) for base fusion generate more details.

4.5 Extension: Medical Image Fusion Experiments

In this section, we extend DCINN to a typical image fusion task (MIF), namely an MRI-CT image fusion task. We collect 360 pairs of MRI and CT images from the Harvard medical image dataset,[19] in which 340 pairs are used for training and

[18] The reasons we chose IFCNN rather than U2Fusion or YDTR are that IFCNN yields better visual quality than U2Fusion and YDTR tends to generate artifacts.

[19] http://www.med.harvard.edu/AANLIB/home.htm.



CT

MRI

U2Fusion

SwinFusion

IFCNN

DCINN (Max+Mean)

DCINN (IFCNN+Mean)

Fig. 18 Visual comparisons on the Harvard medical image dataset

20 pairs are used for testing. For a fair comparison, we employ the same loss function as SwinFusion (Ma et al., 2022) to train our DCINN, and the hyperparameters involved in the loss function are also set to SwinFusion (Ma et al., 2022). We compare the proposed DCINN with SwinFusion (Ma et al., 2022), U2Fusion (Xu et al., 2022), and IFCNN (Zhang et al., 2020) for this medical image fusion application. Moreover, we also report the experimental results of DCINN with different base fusion rules.[20] DCINN with the different fusion rules are named "DCINN (base fusion rule + detail fusion rule)", i.e., "DCINN (Max+Mean)" denotes the DCINN with the max rule as base fusion rule and mean rule as detail fusion rule in Table 11 and Fig. 18. According to Table 11, the proposed DCINN that uses IFCNN (Zhang et al., 2020) as base fusion rule outperforms all the competitors on all the metrics. Besides, the proposed DCINN that uses the max fusion rule performs very well against the three compared methods, i.e., IFCNN (Zhang et al., 2020), U2Fusion (Xu et al., 2022), and SwinFusion (Ma et al., 2022), and slightly worse than our DCINN using IFCNN as base fusion rule. In addition, according to Fig. 18, both the versions of our DCINN achieve promising visual results, especially for the proposed DCINN with IFCNN (Zhang et al., 2020) as fusion rule.

5 Conclusion

In this paper, we observed that LIM-based approaches, LDM-based methods, and INN have complementary advantages and disadvantages when dealing with image fusion challenges. Based on our findings, we proposed the DCINN paradigm to capitalize on their advantages while avoiding their disadvantages. The given DCINN has three core components: a decomposition module that transforms the source images into detail and base components that allow learning detail mapping; an ANet that extracts abundant auxiliary features directly from the source images; and a CINN module that is conditioned on the auxiliary features to learn the detail mapping. For the pansharpening, HMF, and IVF tasks, we compared the proposed DCINN and some SOTA approaches. Extensive experiments showed that the DCINN method achieves superior quantitative performance and visual quality. Moreover, during cross-dataset evaluation, the proposed DCINN exhibited very competitive generalization performance on the pansharpening task. In addition, we conducted thorough ablation studies to validate the effectiveness of our novel design. The experimental analysis demonstrated that: (a) using the ANet to extract auxiliary features helps CINN to better learn the detail mapping providing global features; (b) compared to widely-used DL architectures, the CINN can better preserve image details; and (c) using the INN to learn image mapping can result in degraded performance due to the INN low capacity. Furthermore, our DCINN is capable of solving multi-focus image fusion and multi-exposure image fusion challenges, which will represent our future research topics exploiting the proposed paradigm.

Acknowledgements This research is supported by National Natural Science Foundation of China (12271083, 12171072), Natural Science Foundation of Sichuan Province (2022NSFSC0501), and National Key Research and Development Program of China (Grant No. 2020YFA0714001).

[20] We simply adopt the mean rule as detail fusion rule.

References

Adu, J. J., Gan, J. H., Wang, Y., & Huang, J. (2013). Image fusion based on non-subsampled contourlet transform for infrared and visible light image. *Infrared Physics and Technology, 61*, 94–100.

Aiazzi, B., Alparone, L., Baronti, S., & Garzelli, A. (2002). Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing, 40*, 2300–2312.

Alexander, T. (2017). The TNO multiband image data collection. *Data in brief, 15*, 249–251.

Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., & Selva, M. (2008). Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering and Remote Sensing, 74*(2), 193–200.

Andrea, G., Filippo, N., & Luca, C. (2007). Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing, 46*(1), 228–236.

Ardizzone, L., Lüth, C., Kruse, J., Rother, C., & Köthe, U. (2019). Guided image generation with conditional invertible neural networks. CoRR.

Barata, J., & Hussein, M. (2012). The Moore–Penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics, 42*, 146–165.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., & Jacobsen, J. H.(2019). Invertible residual networks. In *International Conference on Machine Learning (ICML)* (pp. 573–582).

Chakrabarti, A., & Zickler, T. (2011). Statistics of real-world hyperspectral images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 193–200).

Choi, J., Yu, K., & Kim, Y. (2010). A new adaptive component substitution based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing, 49*(1), 295–309.

Craig, L., & Bernard, B. (2000). Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. US Patent 6,011,875

Cui, J., Zhou, L., Li F, & Zha, Y. (2022). Visible and infrared image fusion by invertible neural network. In *China Conference on Command and Control (CICC)* (pp. 133–145).

Deng, L. J., Vivone, G., Jin, C., & Chanussot, J. (2021). Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing, 59*(8), 6995–7010.

Deng, L. J., Vivone, G., Paoletti, M. E., Scarpa, G., He, J., Zhang, Y., Chanussot, J., & Plaza, A. (2022). Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine, 10*(3), 279–315. https://doi.org/10.1109/MGRS.2022.3187652

Deng, S. Q., Deng, L. J., Wu, X., Ran, R., Hong, D., & Vivone, G. (2023). PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing, 61*, 1–15. https://doi.org/10.1109/TGRS.2023.3244750

Dian, R. W., Li, S. T., Guo, A. J., & Fang, L. (2018). Deep hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems, 29*(11), 5345–5355.

Dinh L, Krueger D, Bengio Y (2015) Nice: Non-linear independent components estimation. In *Conference on Learning Representations (ICLR) Workshop Track.*

Dong, W. S., Zhou, C., Wu, F. F., Wu, J., Shi, G., & Li, X. (2021). Model-guided deep hyperspectral image super-resolution. *IEEE Transactions on Image Processing, 30*, 5754–5768.

Emiel H, Victor GS, Jakub T, & Welling, M. (2020) The Convolution Exponential and Generalized Sylvester Flows. In *Conference on Neural Information Processing Systems (NeurIPS)* (pp. 18249–18260).

Eskicioglu, A., & Fisher, P. (1995). Image quality measures and their performance. *IEEE Transactions on Communications, 43*(12), 2959–2965.

Fu, X. Y., Wang, W., Huang, Y., Ding, X., & Paisley, J. (2020). Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems, 32*(5), 2090–2104.

Garzelli, A., & Nencini, F. (2009). Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters, 6*(4), 662–665.

Giuseppe, M., Davide, C., Luisa, V., & Scarpa, G. (2016). Pansharpening by convolutional neural networks. *Remote Sensing, 8*(7), 594.

Gomez, A. N., Ren, M., Urtasun, R., & Grosse, R. B. (2017) The reversible residual network: Backpropagation without storing activations. In *Conference on Neural Information Processing Systems (NeurIPS).*

Guan, P. Y., & Lam, E. Y. (2021). Multistage dual-attention guided fusion network for hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–14.

Guo, A. J., Dian, R. W., & Li, S. T. (2023). A deep framework for hyperspectral image fusion between different satellites. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(7), 7939–7954.

Guo, P. H., Zhuang, P. X., & Guo, Y. C. (2020). Bayesian pan-sharpening with multiorder gradient-based deep network constraints. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 950–962.

He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

He, L., Rao, Y. Z., Li, J., Chanussot, J., Plaza, A., Zhu, J., & Li, B. (2019). Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(4), 1188–1204.

Hou, R. C., Zhou, D. M., Nie, R. C., Liu, D., Xiong, L., Guo, Y., & Yu, C. (2020). VIF-Net: An unsupervised framework for infrared and visible image fusion. *IEEE Transactions on Computational Imaging, 6*, 640–651.

Hu, J. F., Huang, T. Z., Deng, L. J., Dou, H. X., Hong, D., & Vivone, G. (2022). Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geoscience and Remote Sensing Letters, 19*, 1–5.

Hu, J. F., Huang, T. Z., Deng, L. J., Jiang, T. X., Vivone, G., & Chanussot, J. (2022). Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 33*(12), 7251–7265.

Huang G, Liu Z, Maaten LVD, & Weinberger, K. Q. (2017) Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (pp. 4700–4708).

Huang, J. J., & Dragotti, P. L. (2022). WINNet: Wavelet-inspired invertible network for image denoising. *IEEE Transactions on Image Processing, 31*, 4377–4392.

Huang, T., Dong, W. S., Wu, J. J., Li, L., Li, X., & Shi, G. (2022). Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Transactions on Computational Imaging, 8*, 201–214.

Jin, C., Deng, L. J., Huang, T. Z., & Vivone, G. (2022). Laplacian pyramid networks: A new approach for multispectral pansharpening. *Information Fusion, 78*, 158–170.

Jin ZR, Zhang TJ, Jiang TX, Vivone, G., & Deng, L. J. (2022b) LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *AAAI Conference on Artificial Intelligence (AAAI)* (pp. 1113–1121).

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference On Learning Representations (ICLR)* (p. 80).

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Lanaras, C., Baltsavias, E., & Schindler, K. (2015). Hyperspectral super-resolution by coupled spectral unmixing. In *International Conference on Computer Vision (ICCV)* (pp. 3586–3594).

Li, H., & Wu, X. J. (2019). DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing, 28*(5), 2614–2623.

Li, H., Wu, X. J., & Kittler, J. (2021). RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion, 73*, 72–86.

Li, H., Xu, T., Wu, X. J., Lu, J., & Kittler, J. (2023). LRRNet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(9), 11040–11052. https://doi.org/10.1109/TPAMI.2023.3268209

Liu, J. G. (2002). Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing, 23*(3), 593–597.

Liu, J. Y., Dian, R. W., Li, S. T., & Liu, H. (2023). SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. *Information Fusion, 91*, 205–214.

Liu, R. S., Liu, J. Y., Jiang, Z. Y., Fan, X., & Luo, Z. (2020). A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Transactions on Image Processing, 30*, 1261–1274.

Liu, X. Y., Liu, Q. J., & Wang, Y. H. (2020). Remote sensing image fusion based on two-stream fusion network. *Information Fusion, 55*, 1–15.

Lu, S. P., Wang, R., Zhong, T., & Rosin, P. L. (2021) Large-capacity image steganography based on invertible neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10816–10825).

Ma, J. Y., Chen, C., Li, C., & Huang, J. (2016). Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion, 31*, 100–109.

Ma, J. Y., Yu, W., Liang, P. W., Li, C., & Jiang, J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion, 48*, 11–26.

Ma, J. Y., Liang, P. W., Yu, W., Chen, C., Guo, X., Wu, J., & Jiang, J. (2020). Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion, 54*, 85–98.

Ma, J. Y., Yu, W., Chen, C., Liang, P., Guo, X., & Jiang, J. (2020). Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion, 62*, 110–120.

Ma, J. Y., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica, 9*(7), 1200–1217.

Ma, Q., Jiang, J. J., Liu, X. M., & Ma, J. (2023). Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution. *Information Fusion*. https://doi.org/10.1016/j.inffus.2023.101907

Miguel, S., Bioucas-Dias, J., Almeida, L. B., & Chanussot, J. (2015). A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing, 53*(6), 3373–3388.

Naoto, Y., Takehisa, Y., & Akira, I. (2012). Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing, 50*(2), 528–537.

Qi, W., Nicolas, D., & Jean-Yves, T. (2015). Fast fusion of multi-band images based on solving a Sylvester equation. *IEEE Transactions on Image Processing, 24*(11), 4109–4121.

Qu, G. H., Zhang, D. L., & Yan, P. F. (2002). Information measure for performance of image fusion. *Electronics Letters, 38*, 1–7. https://doi.org/10.1049/el:20020212

Ran, R., Deng, L. J., Jiang, T. X., Hu, J. F., Chanussot, J., & Vivone, G. (2023). GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution. *IEEE Transactions on Cybernetics*. https://doi.org/10.1109/TCYB.2023.3238200

Rao, Y. J. (1997). In-fibre Bragg grating sensors. *Measurement science and technology, 8*, 355–358.

Tang, W., He, F. Z., & Liu, Y. (2022). YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2022.3192661

Vivone, G., Restaino, R., & Chanussot, J. (2018). Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing, 27*(7), 3418–3431.

Wald, L. (2002). *Data Fusion. Definitions and Architectures—Fusion of Images of Different Spatial Resolutions*. Presses des MINES.

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*(6), 691–699.

Wang, L. G., Guo, Y. L., Dong, X. Y., Wang, Y., Ying, X., Lin, Z., & An, W. (2022). Exploring fine-grained sparsity in convolutional neural networks for efficient inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4474–4493.

Wang W, Zeng WH, Huang Y, Ding, X., & Paisley, J. (2019). Deep blind hyperspectral image fusion. In *International Conference on Computer Vision (ICCV)* (pp. 4150–4159).

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers (ACSSC)* (pp. 1398–1402).

Wesley, R., van Aardt, J., & Fethi, A. (2008). Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing, 2*, 1–28.

Wu, Z. C., Huang, T. Z., Deng, L. J., Huang, J., Chanussot, J., & Vivone, G. (2023). LRTCFPan: Low-rank tensor completion based framework for pansharpening. *IEEE Transactions on Image Processing, 32*, 1640–1655.

Xiao, J. J., Li, J., Yuan, Q. Q., & Zhang, L. (2022). A dual-UNet with multistage details injection for hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–13.

Xiao M, Zheng S, Liu C, Wang, Y., He, D., Ke, G., Bian, J., Lin, Z., & Liu, T. Y. (2020). Invertible image rescaling. In *European Conference on Computer Vision (ECCV)* (pp. 126–144).

Xu, H., Ma, J., Le, Z., Jiang, J., & Guo, X. (2020). FusionDN: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (pp. 12484–12491).

Xu, H., Ma, J. Y., Jiang, J. J., Guo, X., & Ling, H. (2022). U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 502–518.

Xu, Q. Z., Zhang, Y., Li, B., & Ding, L. (2014). Pansharpening using regression of classified MS and pan images to reduce color distortion. *IEEE Geoscience and Remote Sensing Letters, 12*(1), 28–32.

Xu, S., Zhang, J., Zhao, Z., Sun, K., Liu, J., & Zhang, C. (2021). Deep gradient projection networks for pan-sharpening. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1366–1375).

Xu, Y., & Zhang, J. (2021). Invertible resampling-based layered image compression. In *2021 Data Compression Conference (DCC)* (pp. 380–380).

Yan, Y. S., Liu, J. M., Xu, S., Wang, Y., & Cao, X. (2022). MD^{3}Net: Integrating model-driven and data-driven approaches for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–16.

Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., & Paisley, J. (2017). PanNet: A deep network architecture for pan-sharpening. In *International Conference on Computer Vision (ICCV)* (pp. 5449–5457).

Yang, Y., Lu, H. Y., Huang, S. Y., & Tu, W. (2020). Pansharpening based on joint-guided detail extraction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 389–401.

Yang, Y., Wu, L., Huang, S. Y., Wan, W., Tu, W., & Lu, H. (2020). Multiband remote sensing image pansharpening based on dual-injection model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 1888–1904.

Yuhas, R. H., Goetz, A. F., & Boardman, J. W. (1992). Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *Annual JPL Airborne Geoscience Workshop*.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5728–5739).

Zhang, T. J., Deng, L. J., Huang, T. Z., Chanussot, J., & Vivone, G. (2022). A triple-double convolutional neural network for panchromatic sharpening. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2022.3155655

Zhang, X. T., Huang, W., Wang, Q., & Li, X. (2021). SSR-NET: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing, 59*(7), 5953–5965.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., & Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion, 54*, 99–118.

Zhao, R., Liu, T. S., Xiao, J., Lun, D. P., & Lam, K. M. (2021). Invertible image decolorization. *IEEE Transactions on Image Processing, 30*, 6081–6095.

Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., & Zhang, J. (2020). DIDFuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 970–976).

Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., & Van Gool, L. (2023). CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5906–5916).

Zhou, M., Fu, X. Y., Huang, J., Zhao, F., & Hong, D. (2022). Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–14. https://doi.org/10.1109/TGRS.2022.3199210

Zhou, M., Yan, K. Y., Pan, J. S., Ren, W., Xie, Q., & Cao, X. (2023). Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision, 131*(1), 215–242.

Zhou, Z. Q., Wang, B., Li, S., & Dong, M. (2016). Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Information Fusion, 30*, 15–26.

Zhuang, P., Liu, Q., & Ding, X. (2019). Pan-GGF: A probabilistic method for pan-sharpening with gradient domain guided image filtering. *Signal Processing, 156*, 177–190.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.