Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion

Pengwei Liang[1], Junjun Jiang[1⋆], Xianming Liu[1], and Jiayi Ma[2]

[1] Harbin Institute of Technology, Harbin 150001, China.
[2] Wuhan University, Wuhan 430072, China.
erfect2020@gmail.com, {jiangjunjun, csxm}@hit.edu.cn,
jyma2010@gmail.com

Abstract. Image fusion is famous as an alternative solution to generate one high-quality image from multiple images in addition to image restoration from a single degraded image. The essence of image fusion is to integrate complementary information or best parts from source images. The current fusion methods usually need a large number of paired samples or sophisticated loss functions and fusion rules to train the supervised or unsupervised model. In this paper, we propose a powerful image decomposition model for fusion task via the self-supervised representation learning, dubbed **De**composition for **Fusion** (**DeFusion**). Without any paired data or sophisticated loss, DeFusion can decompose the source images into a feature embedding space, where the common and unique features can be separated. Therefore, the image fusion can be achieved within the embedding space through the jointly trained reconstruction (projection) head in the decomposition stage even without any fine-tuning. Thanks to the development of self-supervised learning, we can train the model to learn image decomposition ability with a brute but simple pretext task. The pretrained model allows for learning very effective features that generalize well: the DeFusion is a unified versatile framework that is trained with an image fusion irrelevant dataset and can be directly applied to various image fusion tasks. Extensive experiments demonstrate that the proposed DeFusion can achieve comparable or even better performance compared to state-of-the-art methods (whether supervised or unsupervised) for different image fusion tasks.

Keywords: Image fusion, Self-supervised learning, Image decomposion

1 Introduction

The scene perception is a long-standing goal of machine vision, in which the scene is digitized by multiple hardware sensors. Each sensor can capture only parts of information from the scene at a time due to hardware limitations. In order to represent the scene accurately and effectively, image fusion is pushed forward

⋆ Corresponding author (jiangjunjun@hit.edu.cn).

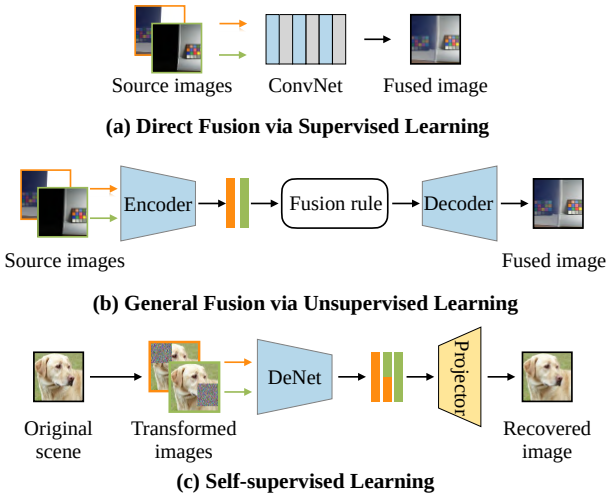

Fig. 1. Paradigms of different image fusion methods. Most existing image fusion methods based on deeplearning can be classified into (a) and (b). We draw the insights from the essence of image fusion and propose a new image fusion framework based on self-supervised learning.

to integrate the complementary features of multiple source views in the same scene, thus generating a high-quality image for the downstream high-level tasks or human perception [43]. For example, the multi-exposure fusion (MEF) utilizes multiple low dynamic range (LDR) images to obtain a single high dynamic range (HDR) image [52,23]; the multi-focus fusion (MFF) combines multiple images with different focus areas into a single all-in-focus image [53]. An essential step in image fusion methods is to effectively represent the source images. In the early years, some classical feature representation and decomposition methods have been introduced into image fusion, such as wavelet [26], pyramid [37], edge-preserving filter [14], sparse coding and dictionary learning [42]. Driving from the signal processing perspective, these manually designed feature representation approaches poorly understand the semantic knowledge of images, which limits the generalizability of those models.

Recently, deep learning has been introduced to address the limitations by adaptively learning image representations from large-scale dataset, and push forward the frontier of image fusion research. In the pioneer works, researchers simply regard the network as an optimizer, which is used to model the relationship between the source images and the target fusion result, and we called the framework 'direct fusion via supervised learning' as shown in Fig. 1a. Obviously, these models exhibit a major flaw: obtaining the paired source images and ground-truth fused image would be difficult [3], if not impossible, in some scenario, *e.g.*, infrared-visible image fusion [18]. An alternative solution is to abandon the supervision information and carefully design some auxiliary losses

to maintain the consistency between the fused image and source images [47]; or leverage tailored fusion rules to perform fusion at the semantic bottleneck layer of some pretrained networks (such as AutoEncoder [33,13]), as shown in Fig. 1b. Although these advancements expand the applicable scenarios, they still suffer from a serious flaw: their performance seriously depends on the human knowledge about the auxiliary loss and fusion rule.

To address the aforementioned issues, we propose a self-supervised learning framework for image fusion, dubbed **DeFusion**, without needing sophisticated loss functions or fusion rules shown in Fig. 1c. We can learn from the definition of image fusion that the essence of image fusion is to integrate the complementary information of multiple source images. Therefore, if we can decompose the source images into the unique component and the shared common component of all images, the target fusion image can be generated by simply combining the components. The remaining question is: *how to decompose the source images to obtain the unique and common components without any supervision*?

Given the source images, it is very hard for us to obtain the supervision information to guide the prediction of the unique and common components. In this paper, we design a pretext task, called common and unique decomposition (CUD), to perform image decomposition under a self-supervised learning framework. We are dedicated to decomposing the multiple source images into unique and common feature representations to accomplish the unsupervised image fusion (*i.e.*, *fusion from decomposition*). As shown in Fig. 2, we design a specific image augmentation strategy that will replace some patches of the original scene $\boldsymbol{x}$ with noise to generate two 'source images', $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$. Afterwards, they are fed into the decomposition network DeNet to get the common features f_c, and the unique features f_u^1 and f_u^2 corresponding to $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$. Acquiring the embedding features, we then apply two projection heads, the common projection head P_c and the unique projection head P_u, to produce the common and unique images (parts) of the source images $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$. Under the specific image augmentation strategy, we can easily generate the supervision of the projected common and unique images. In addition, the combined features f_c, f_u^1, f_u^2 are also fed into a reconstruction projection head P_r to reconstruct the original scene $\boldsymbol{x}$. In the inference phase, we can decompose the source images into common and unique semantic representations and reconstruct the fused images from the combined features, as shown in Fig. 3. In this way, the combination of decomposed common and unique features provides explainable information for fused images and bypasses the difficulties of developing sophisticated loss functions or fusion rules.

In summary, we can summarize our contributions as follows: (i) We propose a novel image fusion method called DeFusion by decomposing the source image based on a self-supervised learning framework. (ii) We design a pretext task, called CUD, for image fusion, which does not rely on the existing supervised image fusion dataset, sophisticated loss functions and fusion rules. (iii) The proposed DeFusion is trained only with the COCO dataset and can be used as a unified and versatile framework for various image fusion tasks without any further fine-tuning or introducing an additional fusion rule. It achieves compa-

rable or even better performance compared to the most competitive image fusion methods (including supervised ones) on various types of fusion tasks.

2 Related Work

2.1 Deep Learning-based Image Fusion

In the past decades, image fusion based on deep learning has gained much spotlight in research community. Liu *et al.*[17] first trained a binary classification convolutional neural network for the multi-focus image fusion task. Inspired by this, many multi-focus image fusion methods via supervised-learning had been proposed [54,9,40]. Specifically, Zhang *et al.*[54] proposed an end-to-end fusion method called IFCNN that used RBG image and the corresponding depth image to simulate training samples. However, these methods are hard to transfer into the multi-modal image fusion task, *e.g.*, infrared-visual image fusion, in which ground truth does not naturally exist.

Rather than simulating training data from ground truths, unsupervised methods focus on designing fusion rules and loss functions [33,13,20,38,19]. Typically, the DeepFuse [33], DenseFuse [13] employed the fusion rule (addition strategy) into the extracted features on the bottleneck of autoencoder. The U2fusion [38], MEFNet [21], PMGI [47] designed multiple losses with considerable variation for the same fusion task. However, the design of sophisticated losses requires the human knowledge, which limits its generalizability. In this work, we decompose the multiple images into semantic embeddings via self-supervised learning, and thus avoiding the design of fusion rules and sophisticated losses.

2.2 Image Decomposition Model

In traditional image fusion methods, the image decomposition model is one of mainstream fusion strategy. A typical option of image decomposition is to use a set of predefined basis functions, *e.g.*, wavelets [26], conventional pyramids [2,32], to represent the images. In addition, the average filter is employed to decompose the images into base layer and detail layer [14]. Nevertheless, the decomposed components still rely on manually tailored fusion rules to extract the useful information, in which the distortion information may incorrectly be retained into fused results [39]. Different with the traditional image decomposition model, the CU-Net [6] used a coupled dictionary learning algorithm to jointly decompose multiple source inputs into necessary features and avoid designing the fusion rules. The DRF [39] decomposed the visible and infrared images into scene and sensor modality representations to alleviate the disadvantage of fusion with fusion rules. However, since the supervision of decomposed components is insufficient during training, these methods may fall into a trivial solution that the decomposed components may be meaningless. Instead of regarding the decomposition components as intermediate procedures (byproducts) and only focusing on the prediction results, we focus on decomposing the multiple source images into unique and common feature representations to assist the image fusion task.

2.3 Self-Supervised Learning

Self-supervised learning is a paradigm to obtain the useful representations from large unlabeled data [16]. Practically, useful representations are extracted by specific pretext task [8]. The pretext tasks are designed to solve the complementary prediction task where we remove part inherent attributes of the image (*e.g.*, the color) to recover it. Recently, a large body of novel pretext tasks had been proposed and had made great progress. One class of pretext is to exploit the geometric transformations of image, such as solving jigsaw puzzles [27,29], recognizing image orientation [12], learning to count [28], image colorization [49]. The others is to leverage multi-modal information (*e.g.*, predicting depth from RGB [34,50], detecting misalignment between audio and visual streams [30]).

3 Method

In this section, we first introduce the self-supervised learning pipeline for image fusion in Section 3.1. Next, we elaborate on a carefully designed pretext task (*i.e.*, CUD) for self-supervised image decomposition and fusion in Section 3.2. Finally, the implementation details are presented in Section 3.3.

3.1 Self-supervised Learning for Image Fusion

Self-supervised learning pipeline. Suppose that we are given an unlabeled image dataset D. For each image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$ in D, we apply a random data augmentation from a set of image transformations $\mathcal{T}$ into $\boldsymbol{x}$ to generate the distorted views $\boldsymbol{x}^i$. The distorted view will be fed into a convolution network to obtain corresponding embeddings. To generate a powerful embedding representation, the convolution network needs to be trained in solving pretext tasks, such as predicting image rotations [12], image colorization [49], and jigsaw puzzles [24]. After pre-training by the pretext task, the embedding representation can be used for downstream tasks.

Image fusion by self-supervised learning. According to the type of sensor that obtains the source images, we can further classify the image fusion into single-modal fusion and multi-modal fusion. For single modality fusion, the observed images are generated from the same type of sensor but with different settings. For multi-modal fusion, the source images come from different types of sensors with different imaging mechanisms, such as infrared-visible. Although the source images exhibit obvious discrepancies, whether the single-modal or the multi-modal cases, they all are transformed from the same scene and represent different (complementary) parts of the scene. Moreover, the goal of image fusion is to retain the vivid information of the multiple inputs to generate the fused image. The procedure of $original\ scene \twoheadrightarrow source\ images \rightrightarrows fused\ image$ is similar to the pipeline of embedding representation in self-supervised learning. Therefore, similar to the self-supervised learning pipeline, we assign the source images to represent the distorted views which will be fed into $\phi_\theta(\cdot)$ to extract the

embeddings, and then apply the embeddings to produce the final fused image via a projection head. In the following, we will present how to practice self-supervised learning of the procedure *original scene* $\twoheadrightarrow$ *source images* $\rightrightarrows$ *fused image*.

3.2 Details of CUD Pretext Task

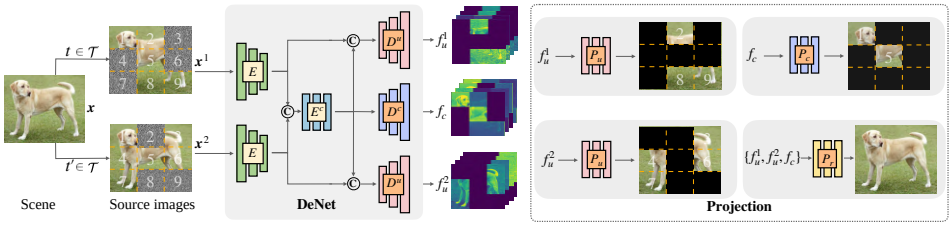

Fig. 2. An overview of our proposed self-supervised image decomposition and fusion method.

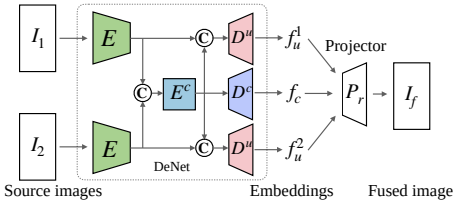

Fig. 3. The testing pipeline of the DeFusion method.

In the typical self-supervised learning paradigm, the learned embeddings have strong representation ability by training with some pretext tasks and can be used for downstream tasks by fine-tuning with limited supervision. However, as for the image fusion task, in some cases there is not always supervisory information available. Therefore, we hope that the fusion result can be obtained after the pre-training, without the need for additional supervision information to fine-tune.

Motivated by these observations, we carefully design a specific pretext task, common and unique decomposition (CUD), for image decomposition and image fusion. The CUD task follows a generally acknowledged definition of data fusion to simulate the fusion process, where the goal of image fusion is to combine complementary information from different source images into a synthetic image. For each source image, it shares parts of the scene information with other source images while retaining some unique information. Therefore, the CUD pretext task will force each source image to be decomposed into two parts: the unique

features and the common features. After pre-training, the obtained common and unique embeddings can be directly used for the image fusion task.

As discussed in Section 3.1, the unlabeled image $\boldsymbol{x}$ corresponds to the original scene in image fusion. Note that we conjecture that the scene in the image fusion involves the most comprehensive information while each observed degraded views $\boldsymbol{x}^i$ only can reflect part of original scene. In CUD, we use a random mask M_i and Gaussian noise n to simulate the degraded transformation $\mathcal{T}$:

$$\boldsymbol{x}^i = M_i(\boldsymbol{x}) + \bar{M}_i(n), \tag{1}$$

where $\bar{M}_i$ is the logical negation operator of mask M_i. To simplify notations, we only focus on the case where the number of source images is 2:

$$\begin{gathered}\boldsymbol{x}^1 = M_1(\boldsymbol{x}) + \bar{M}_1(n), \boldsymbol{x}^2 = M_2(\boldsymbol{x}) + \bar{M}_2(n), \\ s.t. \quad M_1 + M_2 \succ 0.\end{gathered}$$

The constraint is used to ensure that all information in the original scene is included in the augmented images. Different from the traditional inpainting-based pretext tasks [31,7], which remove the remaining regions, here we fill the remaining with random noise, and this will guarantee that the unique information of one image is independent with the counterpart of the other image.

We show a simple example of the transformed images in Fig. 2. The simulated images $\boldsymbol{x}^1, \boldsymbol{x}^2$ are fed into the DeNet $\phi_\theta(\cdot)$ to generate the embedding:

$$f_c, f_u^1, f_u^2 = \phi_\theta(\boldsymbol{x}^1, \boldsymbol{x}^2), \tag{2}$$

where f_c denotes the common embedding of source images, f_u^1, and f_u^2 represent the unique embeddings of the $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$, respectively. Similar to the self-supervised learning pipeline [5], we also introduce some projection heads to project the embeddings into the image space. For the common embedding f_c, the projection $\hat{\boldsymbol{x}}_c = P_c(f_c)$ in the image space should be close to the intersection regions of source images $\boldsymbol{x}_c = M_1(\boldsymbol{x}) \cap M_2(\boldsymbol{x})$. In a similar vein, $\boldsymbol{x}_u^1 = M_1(\boldsymbol{x}) \cap \bar{M}_2(\boldsymbol{x})$ and $\boldsymbol{x}_u^2 = \bar{M}_1(\boldsymbol{x}) \cap M_2(\boldsymbol{x})$ are the ground truths corresponding to the projection of embeddings, $P_u(f_u^1)$ and $P_u(f_u^2)$, respectively. Since the embeddings have encoded the whole semantic information of scene, the projection of embeddings $P_r(f_c, f_u^1, f_u^2)$ should be corresponding to the original scene $\boldsymbol{x}$. As a consequence, our loss function computes the mean absolute error (MAE) between the four projected results and the corresponding original images or masked regions in the pixel space.

3.3 Implementation Details

Details of the network. The overall network $\phi_\theta(\cdot)$ is similar to the bottleneck structure, which can prevent a trivial identity mapping from being learned. The $\phi_\theta(\cdot)$ can be split into three parts: the encoder $E_\theta(\cdot)$, the ensembler $E_\theta^c(\cdot)$, and the decoder $D_\theta(\cdot) = \{D_\theta^u(\cdot), D_\theta^c(\cdot)\}$. As shown in Fig. 2, the $E_\theta(\cdot)$ includes three maxpool layers and residual layers [45] to obtain the compressive representations

whose feature maps size is $\frac{H}{8} \times \frac{W}{8} \times k$. Subsequently, the representations $E_\theta(\boldsymbol{x}^1)$ and $E_\theta(\boldsymbol{x}^2)$ are jointly fed into the $E^c_\theta(\cdot)$ to extract abstract common representation in which the $E^c_\theta(\cdot)$ is composed of only a residual layer. Afterwards, the $D_\theta(\cdot)$ which is composed of several upsample layers and residual layers is applied to generate the corresponding embeddings with the different outputs of the $E_\theta(\cdot)$ and $E^c_\theta(\cdot)$. For instance, the embeddings f^1_u are extracted by the $D^u_\theta(\cdot)$ with the input $\left[E_\theta(\boldsymbol{x}^1); E^c_\theta[E_\theta(\boldsymbol{x}^1); E_\theta(\boldsymbol{x}^2)]\right]$ where $[;]$ is the concatenation operator; and similarly for the embeddings $f^2_u = D^u_\theta\left[E_\theta(\boldsymbol{x}^2); E^c_\theta[E_\theta(\boldsymbol{x}^1); E_\theta(\boldsymbol{x}^2)]\right]$; For the embedding f_c, it only takes the $E^c_\theta[E_\theta(\boldsymbol{x}^1); E_\theta(\boldsymbol{x}^2)]$ as input. In addition to the convolution network $\phi_\theta(\cdot)$, the projection heads $P_c(\cdot), P_u(\cdot), P_r(\cdot)$ that consist of upsample layers and ResNest layers with learned parameters. More details are shown in the supplementary materials.

Training details. We train the convolution network $\phi_\theta(\cdot)$ and projection heads for the CUD pretext task on scenes from large-scale dataset, *i.e.*, COCO dataset [15]. We select 50k images from it to build up the training dataset. During the training phase, our model is trained with an Adam optimizer [11], 50 epochs, batch size 8, and the initial learning rate is set to $1e$-3 that decreased by half each 10 epochs. As for data augmentation, we randomly reshape and crop the images to 256×256. To better simulate degraded process, the M in Eq. 1 is designed as the combination of two random masks with different resolution.

4 Experiments

In this section, we evaluate the DeFusion on multiple tasks such as multi-exposure image fusion, multi-focus image fusion, and visible infrared image fusion. The qualitative and quantitative experiment results demonstrate that DeFusion achieves comparable or even better performance compared to the state-of-the-art (SoTA) methods. In the next subsection, we only show a few examples for each fusion task, the more quantitative fused results can be found in the supplementary materials.

4.1 Comparisons on Different Fusion Tasks

Multi-Exposure Image Fusion. We compare the DeFusion with six SoTA methods, including unified fusion methods, *i.e.*, CU-Net [6], U2Fusion [38], IFCNN [54], PMGI [47], specific-task fusion methods, *i.e.*, DeepFuse [33], MEFNet [21]. For a fair and comprehensive comparison, we evaluate comparison methods on the most comprehensive MEFB benchmark [52] and the largest SICE dataset [3]. Note that the MEFB benchmark contains 100 image pairs collected from multiple public datasets [44,33,22].

Qualitative results on the MEFB benchmark are reported in Fig. 4, where we highlight two regions in each example. As can be seen, the CU-Net suffers random shadowy artifacts, and the IFCNN shows color distortion across the whole images in the first example. The MEFNet shows poor performance on the fusion of semantic information, resulting in inconsistent background with

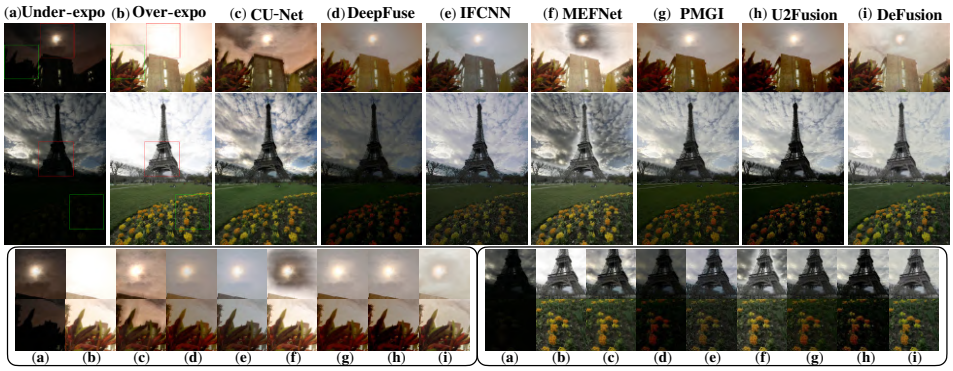

Fig. 4. Qualitative comparison of the DeFusion with 7 SoTA methods on 2 multi-exposure image pairs on the MEFB benchmark.

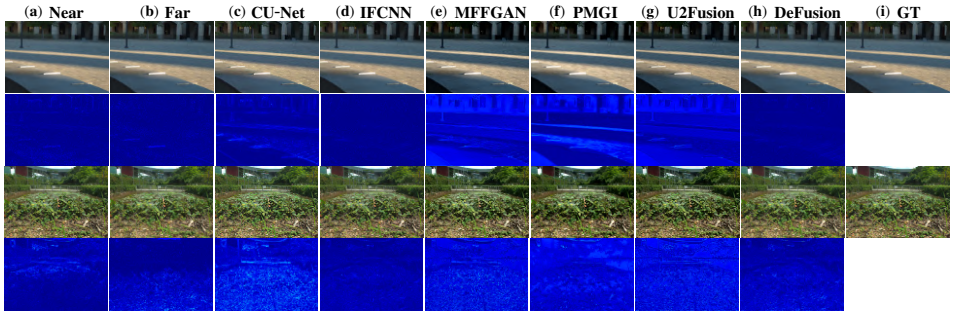

Fig. 5. Qualitative comparisons of multi-focus images fusion results. We provide the enhanced residual maps for each result of comparison and input images to highlight the difference with GT.

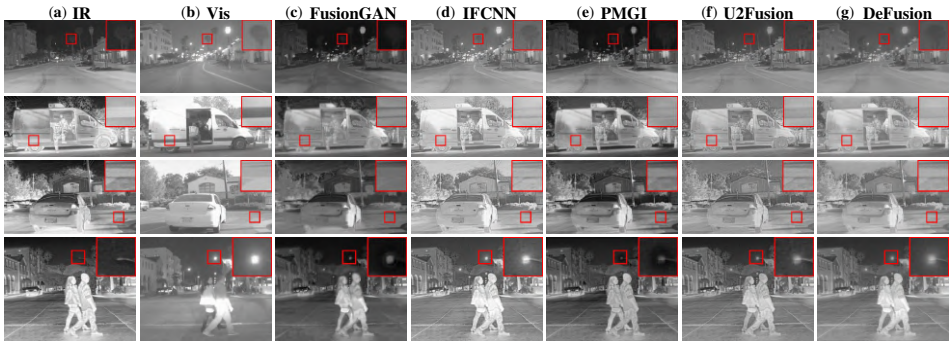

Fig. 6. Qualitative comparisons of visible and infrared image fusion results.

Table 1. Quantitative results on MEFB [52] and SICE [3] datasets for multi-exposure image fusion. The **best**, **second best** and the third best results are marked with red, bold, and underlined.

Method	MEFB [52]						SICE [3]					
	CE	QCV	SSIM	MEF-SSIM	IS	LPIPS	CE	QCV	SSIM	MEF-SSIM	IS	LPIPS
CU-Net	4.800	425.5	0.547	0.794	6.470	0.359	4.728	345.6	0.486	0.742	7.564	0.389
DeepFuse	4.993	363.0	0.544	0.796	6.346	0.380	5.262	189.3	0.523	0.797	**8.391**	0.322
IFCNN	4.943	247.7	0.573	0.818	6.776	0.335	4.551	290.2	0.492	0.697	8.453	0.372
MEFNet	**4.257**	593.4	**0.593**	0.796	6.432	0.321	5.102	505.7	**0.526**	0.711	8.068	0.358
PMGI	4.698	293.9	0.547	**0.822**	6.521	0.336	5.556	294.7	0.480	0.740	7.973	0.375
U2Fusion	4.526	**253.8**	0.526	0.815	**3.438**	**6.745**	**0.332**	209.5	0.488	**0.796**	8.314	**0.346**
DeFusion	2.881	262.3	0.608	0.827	6.587	**0.332**	2.830	**207.7**	0.571	0.788	7.869	0.353

Table 2. Quantitative results on the dataset collected by [53] and the Real-MFF [48] for multi-focus image fusion.

	Method	Dataset[53]		Real-MFF noref [48]		Real-MFF [48]	
		SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
Super-vised	IFCNN	**0.905**	**26.91**	**0.964**	**32.93**	0.983	36.92
Unsup-ervised	CU-Net	0.874	24.88	0.900	26.66	0.938	29.17
	MFFGAN	0.879	24.30	0.811	22.81	0.850	24.14
	PMGI	0.865	20.88	0.890	24.09	0.903	24.66
	U2Fusion	0.815	21.94	0.849	23.99	0.880	25.26
	DeFusion	0.928	28.13	0.969	33.61	**0.971**	**33.88**

Table 3. Quantitative results on TNO [36] and RoadScene [38] datasets for visible-infrared image fusion.

Method	TNO [36]				RoadScene [38]			
	CE	QCV	SSIM	CC	CE	QCV	SSIM	CC
FusionGAN	2.489	954.7	0.631	0.461	1.723	1225.6	0.595	0.561
IFCNN	1.746	340.2	0.701	0.519	0.989	509.8	0.707	0.627
PMGI	1.751	481.0	0.696	0.534	1.277	1024.1	0.668	0.591
U2Fusion	**1.549**	586.1	0.727	0.552	**0.786**	908.2	**0.723**	**0.635**
DeFusion	1.487	**425.3**	**0.715**	**0.539**	0.767	**647.5**	0.727	0.652

heavy halo effects. In addition, the DeepFuse, PMGI and U2Fusion convert the RGB into YCbCr color space and just focus on fusing the Y channel, which may suffer the color shift issue. For example, in the highlighted region of the second sample, the generated flowers of those methods are painted orange, while the original color of the flowers is yellow. The results generated by DeFusion perform are visually pleasant, whose fused objects show a consistent and uniform appearance while avoiding artifacts and distortions. For instance, as shown in the second sample, the DeFusion introduces the details of under-exposed image into fused image while remaining the brightness of over-exposed image rather than the under-exposed image. It demonstrates that the DeFusion can fuse the source images at the semantic feature level by using the embedding representations.

Quantitative comparisons are performed on the MEFB and SICE dataset in Tab 1. We introduce six commonly used metrics, *i.e.*, cross entropy (CE), QCV [4], SSIM, MEF-SSIM [22], IS [35], and LPIPS [51] to measure the quality of fused images. Since the ground truths are unavailable, all metrics are computed by comparing with the two source images as in many previous works. As can be seen, DeFusion ranks first in terms of CE and SSIM on all datasets, and achieves comparable results in terms of QCV, LPIPS and MEF-SSIM.

Multi-Focus Image Fusion. We compare the DeFusion with the following five SoTA methods: CU-Net [6], IFCNN [17], MFFGAN [46], PMGI [47], and U2Fusion [38]. All comparison methods are evaluated on the Real-MFF dataset

[48] and dataset in [53]. To the best of our knowledge, the Real-MFF is the biggest realistic public dataset which provides the realistic source images with corresponding ground truth captured by a light field camera. In addition to the Real-MFF dataset, we also use the collected dataset by Zhang [53], which includes three MFIF datasets, *i.e.*, the Lytro dataset [25], the MFI-WHU dataset [46] and the MFFW dataset [41].

The qualitative results on the Real-MFF are shown in Fig. 5. The results of quantitative comparison on the dataset [53] and Real-MFF [48] are shown in Tab 2. From these reported results, we can learn that the performance of DeFusion goes beyond the other unsupervised methods, and has achieved comparable performance to the IFCNN that is trained via the supervised learning.

Infrared Visible Image Fusion. We compare DeFusion with four SoTA methods: IFCNN [54], FusionGAN [20], PMGI [47], and U2Fusion [38]. For infrared visible image fusion, TNO [36] is a widely used dataset, and RoadScene [38] is a challenging dataset whose infrared images show rich thermal textures. We employ them to explore the performance of comparison methods.

Some qualitative results of the RoadScene dataset are shown in Fig. 6. Due to the physical differences, the source images captured by two different cameras are quite different, which may cause the fusion methods fail to distinguish the object from the background. For example, FusionGAN mixes up the object of visible image and background of infrared image, resulting in the object disappears in the fused result, as shown in the highlighted region of first example. In the second example of Fig. 6, IFCNN, FusionGAN, and PMGI just preserve the edge of stripe and miss the key filled color information in their fused results. A similar phenomenon shows in the third example where the textual information is not well preserved by FusionGAN and IFCNN. In addition, it is of paramount importance for the fusion task to preserve useful information into fused results. However, U2fusion is inclined to preserve excessive infrared information, which may remain some noise of infrared image into the fused image shown in the fourth example. In contrast, our method can well balance these effects and preserve much semantic information.

Quantitative comparisons are shown in Tab. 3 where we use four metrics, *i.e.*, CE, QCV, SSIM, and CC to evaluate all comparison methods. On the RoadScene dataset, the DeFusion ranks first on the CE, SSIM, and CC, indicating that the generated fused results are higher similarity to the source images. For QCV, DeFusion also achieves comparable results. In addition, the results of the TNO dataset show similar performances to those of RoadScene.

4.2 Visualizing Feature Embeddings

In this section, we will demonstrate the unique and common representation ability of our method by some toy and read examples. We take some images from Set5 [1] dataset as the original scenes and apply several image augmentation strategies to them. In principle, the strategies can be classified into toy examples, *i.e.*, the 1st-3rd rows, and real examples, *i.e.*, 4th-5th rows.

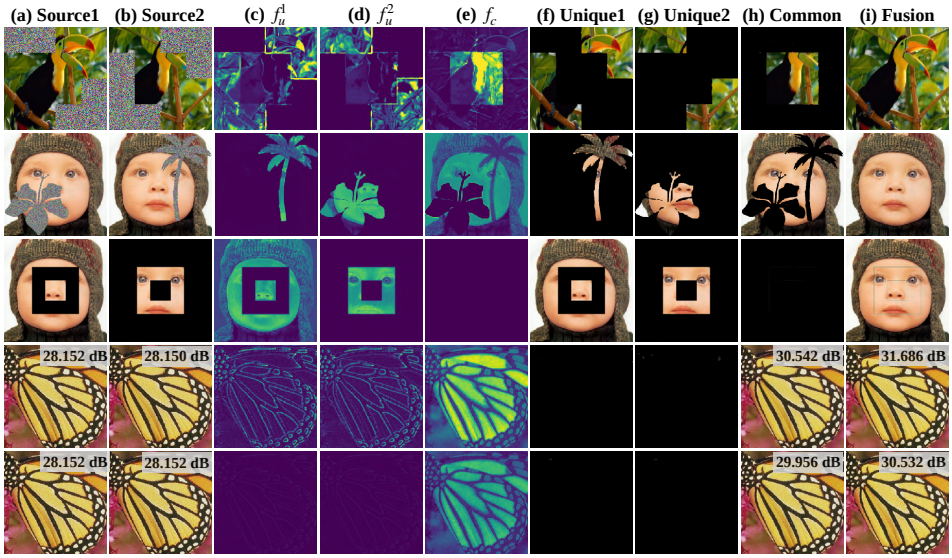

Fig. 7. Visualization of feature embeddings intermediate results for some toy and real examples.

Toy experiments. As shown in Fig. 7, for the first example, the image augmentation is similar to the one defined in Eq. 1. In this sample, the common and unique components are accurately decomposed from the two source images by the pretrained network. To further verify the generalization of the proposed model, we also demonstrate the results with other image augmentation methods that are different with the one in Eq. 1. In the second example, we change the mask shape to an arbitrary shape with increasing difficulty and ignore the constraint in Eq. 3.2 to allow the region of noise to be overlapped. Although more difficult, the decomposition and fusion results do not lose too much information. In addition, we replace the noise with zeros to generate the source images, as shown in the third example. We can see that the pretrained network can also extract appropriate semantic features and project them into image space. Note that the final fused result shows edge information around the mask, and this is due to the information diffusion caused by the convolution. From these toy examples, we can learn that our network pretrained by the CUD pretext has learned the ability to extract the semantic information to some extent.

Some real results. Instead of synthesizing with specific masks, we add the additive white Gaussian noise into the original image twice, which can be seen as the two augmented source images of the original scene, to see whether our model can obtain the common (*i.e.*, the denoised image) and unique components. In the fourth example, we add two different noises ($\sigma = 10$) to the 'butterfly' image to generate two source images. As can be seen, only the common component is projected into the image while the unique components are deactivated. It is

worth noting that both the fused image and common image are denoised images. In the last example, we feed two identical noisy images with $\sigma = 10$ into our network, and the noise of fused image and common image can be also removed. It demonstrates that our network avoids the trivial mapping between the input and output, and is able to adaptively preserve the scene semantic information.

We also visualize the intermediate embedding representation in real image fusion tasks, as shown in Fig. 8. Taking the first multi-exposure sample as an example, the over-exposed image shows abundant details in the room and meaningless brightness on the windows, while the under-exposed image exhibits landscape outside the window and furnishings with lower sharpness energy in the room. After the DeFusion embeds the multi-exposure image pair, we find that the windows regions are not activated in the over-exposed unique embedding, but are activated in the under-exposure unique embedding. It demonstrates that unique embeddings can adaptively distinguish the effective unique information from meaningless image contents. Moreover, in this case, the common embedding is slightly activated at the edges of the window and lamp, indicating that those edges are salient in both images.

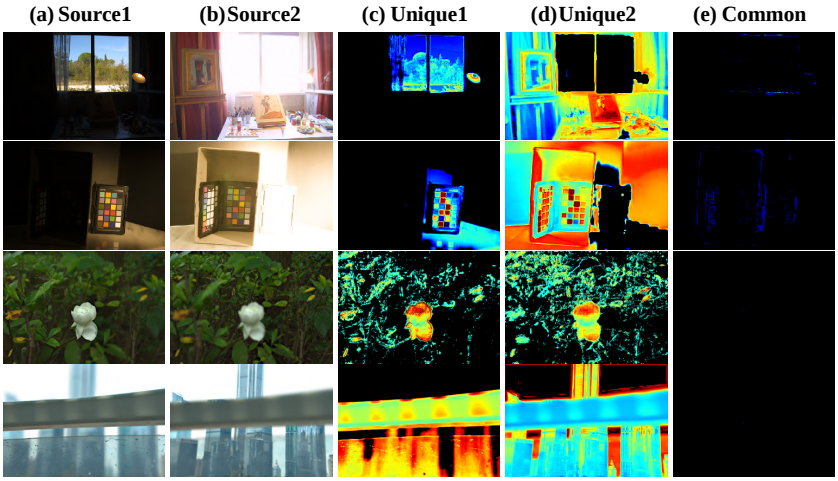

Fig. 8. Feature embedding visualization on multi-exposure image fusion (first two rows) and multi-focus image fusion (last two rows).

In particular, for the multi-focus fusion task, we intuitively infer that the unique useful information should be related with the focus region of the images. However, it is hard to determine which regions should be related to the common information of source images. Interestingly, the statistics of feature representations is consistent with our guess. To vividly describe the statistics, we show a representative example in the second row of Fig. 8. In this case, the unique representations of source images have higher activation values than the corresponding

common representations. Moreover, the focused regions are always corresponding to the activated regions. Note that the common region of multi-focus example shown in Fig. 8(e) is totally black, which may indicate that there are no regions with same amount of defocus in both images.

5 Discussion and Broader Impact

Discussion and limitation. We design an image decomposition model following the essence of image fusion, which is to integrate complementary information from multiple source images and fuse them. Since there are no natural image decomposition components in image fusion, we design a brute but simple pretext task using masks with Gaussian noise to generate the common or unique supervised information. Note that we do not ask the training source images to be strictly aligned with the multiple input images of image fusion task, as our goal is to train the network to learn the ability of decomposing source images into common and unique components. We believe that the obtained decomposed feature embeddings can make the image fusion easier, so that the fused images are generated by a simple convolutional layer called projector, which just likes the last linear layer for classification in general self-supervised learning [10]. Compared to the various pretext tasks for classification in the self-supervised learning, the proposed pretext task for image fusion is **simple and far from a perfect pretext task**. However, the present idea provides a new paradigm for learning multi-source image features jointly, which may provide new directions and considerations for multi-source pre-training. We hope that the new paradigm will inspire more work in the image fusion community.
Broader impact. Recently, the image inpainting-like pretext, *masked autoencoding*, based network pretraining has achieved great success [31,7] in NLP and computer vision. Our DeFusion takes inspirations from these previous works. It can be seen as an extension of these previous single-view masked autoencoding to the multi-view masked autoencoding. Therefore, it provides a paradigm for learning multi-view image features jointly, which may provide new directions and considerations for multi-view pre-training.

6 Conclusion

We present a unified and versatile image fusion framework, *fusion from decomposition*, for multiple image fusion tasks. To obtain an effective representations of the source images, we design pretext task based on the common and unique decomposition (CUD), which can be trained in a self-supervised way and is friendly with our image fusion task. The proposed method achieves comparable or even better performance than previous unsupervised as well as supervised methods. The feature embedding and generalizability of the model have also been verified.

Acknowledgements. The research was supported by the National Natural Science Foundation of China (61971165, 61922027), and also is supported by the Fundamental Research Funds for the Central Universities.

References

1. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of the British Machine Vision Conference. pp. 135.1–135.10. BMVA press (2012) 11
2. Burt, P.J., Kolczynski, R.J.: Enhanced image capture through fusion. In: Proceedings of IEEE International Conference on Computer Vision. pp. 173–182. IEEE (1993) 4
3. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing **27**(4), 2049–2062 (2018) 2, 8, 10
4. Chen, H., Varshney, P.K.: A human perception inspired quality metric for image fusion based on regional information. Information fusion **8**(2), 193–207 (2007) 10
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607. PMLR (2020) 7
6. Deng, X., Dragotti, P.L.: Deep convolutional neural network for multi-modal image restoration and fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 4, 8, 10
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2019) 7, 14
8. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1422–1430 (2015) 5
9. Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., He, K.: FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network. IEEE Transactions on Multimedia **21**(8), 1982–1996 (2019) 4
10. He, K., Chen, X., Xie, S., et al: Masked autoencoders are scalable vision learners. arXiv (2021) 14
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2015) 8
12. Komodakis, N., Gidaris, S.: Unsupervised representation learning by predicting image rotations. In: Proceedings of the International Conference on Learning Representations (2018) 5
13. Li, H., Wu, X.J.: DenseFuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing **28**(5), 2614–2623 (2018) 3, 4
14. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Transactions on Image Processing **22**(7), 2864–2875 (2013) 2, 4
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of European Conference on Computer Vision. pp. 740–755. Springer (2014) 8
16. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering (2021) 5
17. Liu, Y., Chen, X., Peng, H., Wang, Z.: Multi-focus image fusion with a deep convolutional neural network. Information Fusion **36**, 191–207 (2017) 4, 10

18. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. Information Fusion **45**, 153–178 (2019) 2
19. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022) 4
20. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion **48**, 11–26 (2019) 4, 11
21. Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z.: Deep guided learning for fast multi-exposure image fusion. IEEE Transactions on Image Processing **29**, 2808–2819 (2019) 4, 8
22. Ma, K., Zeng, K., Wang, Z.: Perceptual quality assessment for multi-exposure image fusion. IEEE Transactions on Image Processing **24**(11), 3345–3356 (2015) 8, 10
23. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: 15th Pacific Conference on Computer Graphics and Applications (PG'07). pp. 382–390. IEEE (2007) 2
24. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020) 5
25. Nejati, M., Samavi, S., Shirani, S.: Multi-focus image fusion using dictionary-based sparse representation. Information Fusion **25**, 72–84 (2015) 11
26. Nikolov, S., Hill, P., Bull, D., Canagarajah, N.: Wavelets for image fusion. In: Wavelets in signal and image analysis, pp. 213–241. Springer (2001) 2, 4
27. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proceedings of the European Conference on Computer Vision. pp. 69–84. Springer (2016) 5
28. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5898–5906 (2017) 5
29. Noroozi, M., Vinjimoor, A., Favaro, P., Pirsiavash, H.: Boosting self-supervised learning via knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9359–9367 (2018) 5
30. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision. pp. 631–648 (2018) 5
31. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2536–2544 (2016) 7, 14
32. Petrovic, V.S., Xydeas, C.S.: Gradient-based multiresolution image fusion. IEEE Transactions on Image Processing **13**(2), 228–237 (2004) 4
33. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4714–4722 (2017) 3, 4, 8
34. Ren, Z., Lee, Y.J.: Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 762–771 (2018) 5
35. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Proceedings of the Advances in Neural Information Processing Systems **29** (2016) 10

36. Toet, A.: TNO Image Fusion Dataset (4 2014). https://doi.org/10.6084/m9.figshare.1008029.v1, https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029 10, 11
37. Wang, W., Chang, F.: A multi-focus image fusion method based on Laplacian pyramid. J. Comput. **6**(12), 2559–2566 (2011) 2
38. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2Fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 4, 8, 10, 11
39. Xu, H., Wang, X., Ma, J.: DRF: Disentangled representation for visible and infrared image fusion. IEEE Transactions on Instrumentation and Measurement **70**, 1–13 (2021) 4
40. Xu, K., Qin, Z., Wang, G., Zhang, H., Huang, K., Ye, S.: Multi-focus image fusion using fully convolutional two-stream network for visual sensors. KSII Transactions on Internet and Information Systems (TIIS) **12**(5), 2253–2272 (2018) 4
41. Xu, S., Wei, X., Zhang, C., Liu, J., Zhang, J.: MFFW: A new dataset for multi-focus image fusion. arXiv preprint arXiv:2002.04780 (2020) 11
42. Yang, B., Li, S.: Multifocus image fusion and restoration with sparse representation. IEEE Transactions on Instrumentation and Measurement **59**(4), 884–892 (2009) 2
43. Yang, C., Zhang, J.Q., Wang, X.R., Liu, X.: A novel similarity based quality metric for image fusion. Information Fusion **9**(2), 156–160 (2008) 2
44. Zeng, K., Ma, K., Hassen, R., Wang, Z.: Perceptual evaluation of multi-exposure image fusion algorithms. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX). pp. 7–12. IEEE (2014) 8
45. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Muller, J., Manmatha, R., Li, M., Smola, A.: ResNeSt: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020) 7
46. Zhang, H., Le, Z., Shao, Z., Xu, H., Ma, J.: MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Information Fusion **66**, 40–53 (2021) 10, 11
47. Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J.: Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12797–12804 (2020) 3, 4, 8, 10, 11
48. Zhang, J., Liao, Q., Liu, S., Ma, H., Yang, W., Xue, J.H.: Real-MFF: A large realistic multi-focus image dataset with ground truth. Pattern Recognition Letters **138**, 370–377 (2020) 10, 11
49. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Proceedings of the European Conference on Computer Vision. pp. 649–666. Springer (2016) 5
50. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1058–1067 (2017) 5
51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018) 10
52. Zhang, X.: Benchmarking and comparing multi-exposure image fusion algorithms. Information Fusion (2021) 2, 8, 10
53. Zhang, X.: Deep learning-based multi-focus image fusion: A survey and a comparative study. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 2, 10, 11

54. Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: IFCNN: A general image fusion framework based on convolutional neural network. Information Fusion **54**, 99–118 (2020) 4, 8, 11