Visible and Infrared Image Fusion Using Deep Learning

Xingchen Zhang, *Member, IEEE*, and Yiannis Demiris, *Senior Member, IEEE*

***Abstract*—Visible and infrared image fusion (VIF) has attracted a lot of interest in recent years due to its application in many tasks, such as object detection, object tracking, scene segmentation, and crowd counting. In addition to conventional VIF methods, an increasing number of deep learning-based VIF methods have been proposed in the last five years. Different types of methods, such as CNN-based, autoencoder-based, GAN-based, and transformer-based methods, have been proposed. Deep learning-based methods have undoubtedly become dominant methods for the VIF task. However, while much progress has been made, the field will benefit from a systematic review of these deep learning-based methods. In this paper we present a comprehensive review of deep learning-based VIF methods. We discuss motivation, taxonomy, recent development characteristics, datasets, and performance evaluation methods in detail. We also discuss future prospects of the VIF field. This paper can serve as a reference for VIF researchers and those interested in entering this fast-developing field.**

***Index Terms*—Deep learning, image fusion, multimodal fusion, RGB-T, visible-infrared image fusion.**

I. Introduction

VISIBLE and infrared[1] image fusion (VIF) has been an active research topic for many years [1], [2], [3], [4], [5], [6]. This is due to the complementary properties of visible and infrared images which enable VIF to be applied to many applications, such as object tracking [7], [8], [9], [10], [11], object detection [12], [13], [14], [15], and biometric recognition [16], [17]. Specifically, visible images contain rich texture details, while they are sensitive to illumination change. In contrast, infrared images reveal thermal information that are insensitive to illumination change. However, infrared images do not provide enough texture details. The aim of VIF is to combine information from visible and infrared images to generate a fused image, which is more informative and can therefore facilitate downstream applications, as shown in Fig. 1.

Manuscript received 20 October 2022; revised 28 February 2023; accepted 13 March 2023. Date of publication 30 March 2023; date of current version 30 June 2023. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie under Grant 101025274, and in part by a Royal Academy of Engineering Chair in Emerging Technologies. Recommended for acceptance by K. Schindler. *(Corresponding author: Xingchen Zhang.)*

The authors are with the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BX London, U.K. (e-mail: xingchen.zhang@imperial.ac.uk; y.demiris@imperial.ac.uk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3261282, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3261282

[1]In this paper, we mean thermal infrared images when we mention infrared images without specific explanation.

Fig. 1. The benefit of visible and infrared image fusion. The first row shows visible images while the second row shows corresponding infrared images. The third row illustrates the fusion results using the MGFF method [18]. As can be seen, the fused images contain features from both visible and infrared images.

Before deep learning was introduced to VIF, various conventional VIF methods were proposed. Conventional VIF methods can be generally classified into several types according to their corresponding theories, namely multiscale transform-based methods, sparse representation-based methods, subspace-based methods, saliency-based methods, hybrid methods, and other methods [19]. In recent years, researchers have begun to perform VIF with deep learning techniques [20]. To the best of our knowledge, Wu et al. [21] were one of the first researchers to use deep learning in VIF. In their study, they employed a Deep Boltzmann Machine (DBM) to decide how to fuse coefficients of source images. Since then, many deep learning-based VIF methods have been proposed. As shown in Fig. 2, the number of papers on deep learning-based VIF methods published each year increased very quickly from 2018 to 2022. In terms of methods, both supervised [22], [23], [24] and unsupervised VIF methods [25], [26], [27], [28] have been proposed. In addition, various deep learning models, such as CNNs [29], [30], [31], GANs [6], [24], [32], [33] and transformers [34], [35], [36], [37], have been employed to perform VIF. Fig. 3 shows a timeline of the development of deep learning-based VIF methods with key milestones.

However, the lack of a comprehensive review of deep learning-based VIF methods makes it difficult for researchers, especially those who want to enter this field, to appreciate the past, present and future of this field. There are some review papers in the field of VIF [19], [38], [39], [40]. However,

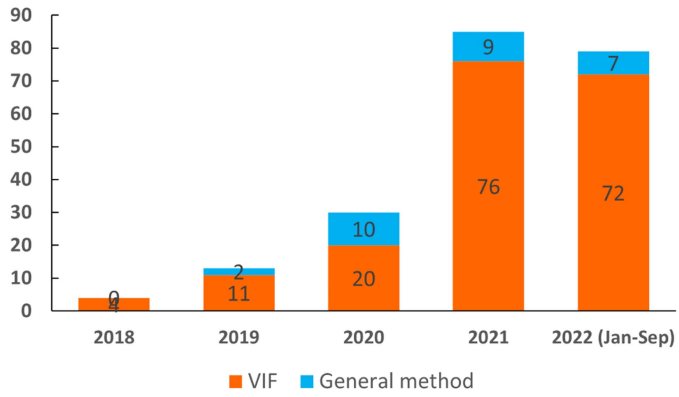


Fig. 2. The number of papers on deep learning-based VIF methods increases quickly. Deep learning-based general image fusion methods that can be applied to several image fusion tasks, including VIF, are also counted.

Ma et al. [19] published their thorough review paper in early 2018, thus it mainly covered conventional VIF methods. Zhang et al. [38] presented a review of VIF-based object tracking. However, that paper focused on object tracking rather than image fusion methods that generate fused images. Sun et al. [39] summarized some deep learning-based VIF methods, but many new deep learning-based methods have been published since then. Zhang et al. [40] did very useful work reviewing deep learning-based image fusion methods, however without focusing on VIF.

In this paper, we present a comprehensive review of deep learning-based VIF methods, covering the motivation, taxonomy, discussions on different types of methods, recent development characteristics, datasets, and future prospects. In summary, the main contributions of this paper lie in the following aspects:

- *A comprehensive review of deep learning-based methods*. We did a thorough screening of the VIF literature and analyzed representative deep learning-based methods. To the best of our knowledge, this is the most comprehensive review of deep learning-based VIF methods to date.
- *A thorough summary of recent development characteristics of deep learning-based VIF methods*. These characteristics are summarized after reading more than 200 deep learning-based VIF papers. Therefore, they can exhibit how deep learning-based VIF methods have evolved in recent years.
- *A detailed discussion on future prospects*. Based on our review and practical experience, future prospects of VIF are discussed. We hope these can help to attract more researchers to work on key issues of the VIF field and can shed some light on this fast-developing field.

The rest of this paper is organized as follows. Section II gives a review of deep learning-based VIF methods. Then, Section III discusses general image fusion methods, followed by the analysis of recent development characteristics of the VIF field in Section IV. Then, Section V summarizes datasets and Section VI discusses performance evaluation methods. Next, Section VII discusses future prospects. Finally, Section VIII concludes this paper.

II. Deep Learning-Based Visible-Infrared Image Fusion Methods

A. Background

Image fusion [122] has been studied for more than 30 years and contains several variants, including visible-infrared image fusion (VIF) [19], multi-focus image fusion (MFIF) [123], multi-exposure image fusion (MEF) [124], medical image fusion (MEDIF) [125], and remote sensing image fusion [126]. In image fusion, VIF is one of the most studied tasks because the complementary properties of visible and infrared images are very beneficial for many applications.

Generally speaking, VIF can be performed at three levels [38]: pixel-level, feature-level, and decision-level. In pixel-level VIF, visible and infrared images are first fused to produce fused images, based on which downstream applications are performed. In feature-level VIF, features are first extracted from both visible and infrared images and then fused. Downstream applications are then performed based on the fused feature. In decision-level VIF, applications are run on visible and infrared images, respectively. The obtained results are then fused to generate the final result. In the image fusion literature, VIF typically means pixel-level VIF methods. *In this paper, VIF refers to pixel-level VIF, i.e., using visible and infrared images to generate fused images.* It is worth mentioning that there is no ground-truth fused image in pixel-level VIF.

B. Motivation for Using Deep Learning in VIF

A VIF method usually consists of three stages, i.e., feature extraction, feature fusion, and image reconstruction, as shown in Fig. 4. First, features of visible and infrared images are extracted. Then, these features are fused. Finally, image reconstruction is performed to obtain the fused image.

For VIF methods, all three stages are essential for good performance. First, complementary features from visible and infrared images should be extracted. To be more specific, detail and texture information in visible images and salient information in infrared images need to be extracted. Second, these complementary features should be fused effectively. Finally, the fused image should be reconstructed effectively. However, these stages are designed manually in conventional VIF methods, which are not very effective in handling various working conditions. Therefore, conventional methods may have limited ability to provide good fused images.

Deep learning has shown great success in various image processing fields. In the past several years, researchers have paid much attention to performing VIF using deep learning, with the aim of achieving improved fusion performance. Note that deep learning may only be applied to a part of the image fusion process.

C. Taxonomy of Deep Learning-Based VIF Methods

Deep learning-based VIF methods can be grouped in different ways. Depending on whether ground truth is used during training, they can be divided into supervised methods and unsupervised methods. Here, supervised means that the method needs

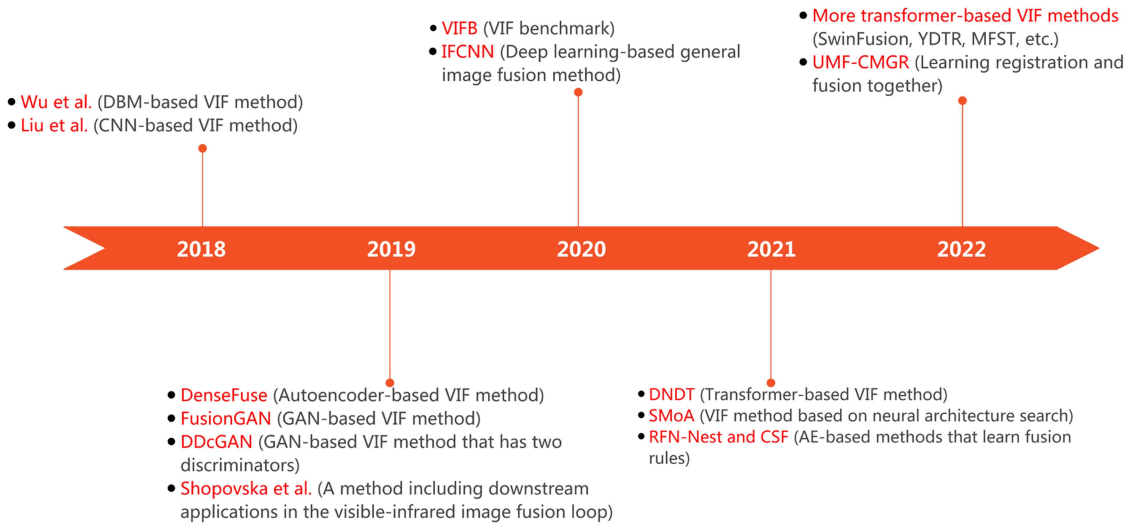


Fig. 3. Development timeline of deep learning-based VIF methods, with some key milestones.

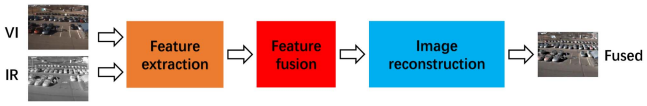


Fig. 4. Three stages of VIF methods, i.e., feature extraction, feature fusion, and image reconstruction.

(a) Single-branch architecture. (b) Dual-branch architecture.

Fig. 5. Single-branch and dual-branch architectures for handling visible and infrared images. In some dual-branch studies, the two deep learning models share weights as denoted by the dashed line.

annotation of some forms, which may not be the ground-truth fused image, in training. According to the type of adopted model, they can be divided into CNN-based methods, autoencoder (AE)-based methods, GAN-based methods, transformer-based methods and others. According to whether the method is end-to-end, VIF methods can be classified as end-to-end methods and non-end-to-end methods. In end-to-end methods, the fused image is generated directly from source images without using manually designed steps, e.g., image decomposition, separate training, and weighted summation based on a weight map. In contrast, at least one manually designed step is required in non-end-to-end methods. Moreover, VIF methods can be grouped into fully convolutional methods and non-fully convolutional methods. Furthermore, different network architectures have been utilized in VIF. In general, there are single-branch and dual-branch architectures, as shown in Fig. 5. In single-branch architectures, visible and infrared images are concatenated in the channel direction to form a 4-channel input [22], [45], which is then fed to a deep learning model. In contrast, in two-branch architectures, visible and infrared images are processed by two branches, respectively. In some cases, the two branches are the same [42], [57], [59], [89] (share weights). However, because visible and infrared images belong to different modalities, in many cases, two separate branches (without weight sharing) are utilized to process them, respectively. An overview of representative deep learning-based VIF method is given in Table I. The following contents of this section are organized according to adopted deep learning models.

D. CNN-Based VIF Methods

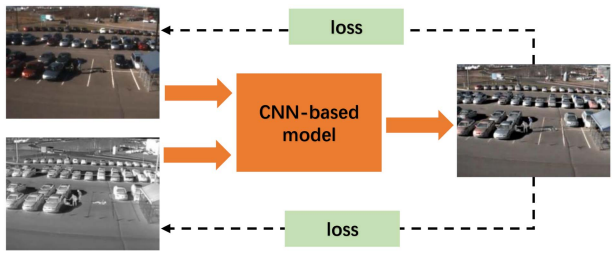
To the best of our knowledge, Liu et al. [29] proposed one of the first CNN-based VIF methods. In that work, a Siamese CNN was used to generate a weight map from source images. The source images were decomposed using Laplacian pyramid, and the weight map was decomposed using Gaussian pyramid. The fusion was then performed in a multiscale manner. The model was trained using high-quality images and their blurred versions generated using multiscale Gaussian filtering and random sampling. As a pioneering deep learning-based VIF method, this method introduced CNN to VIF. Since the work of Liu et al. [29], many CNN-based VIF methods have been proposed.

The main architectures of CNN-based VIF methods are shown in Fig. 6. The main difference between unsupervised and supervised methods lies in how the loss function is constructed. In addition, in some studies, CNN is only applied to a part of the three stages of image fusion. Therefore, the CNN-based model shown in Fig. 6 may contain some other manual steps, e.g., manual fusion rule.

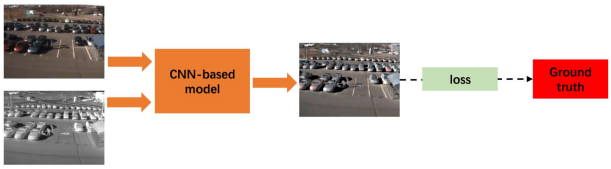
1) Unsupervised Methods: This section introduces representative unsupervised CNN-based VIF methods whose overall architecture is shown in Fig. 6(a). Due to the lack of ground truth, the loss function is usually defined using the fused image and source images. Specifically, the loss function usually contains term(s) constructed from image fusion evaluation metrics.

TABLE I
AN OVERVIEW OF REPRESENTATIVE DEEP LEARNING-BASED VIF METHODS. ONLY PEER-REVIEWED METHODS ARE SUMMARIZED IN THIS TABLE

Name/Reference	Year	Category	General/VIF	Single-branch	End-to-end		Fully convolutional		Supervised		Code
					Yes	No	Yes	No	Yes	No	
Wu et al. [21]	2018	**DBM**	VIF		√			√			
Liu et al. [29]	2018	**CNN**	VIF			√		√	√ (synthetic)		√ (Matlab)
DLF [30]	2018	CNN	VIF		√		√			√ (transfer)	√ (Matlab)
Ren et al. [41]	2018	CNN	VIF		√		√			√ (transfer)	
DenseFuse [42]	2019	**AE**	VIF			√	√			√	√ (Tensorflow)
ResNet [43]	2019	CNN	VIF			√	√			√ (transfer)	√ (Matlab)
Shopovska et al. [44]	2019	CNN	VIF		√		√		√ (application)		
Cui et al. [45]	2019	CNN	VIF	√	√			√		√	
Liu et al. [46]	2019	CNN	VIF		√		√			√	
Wang et al. [47]	2019	CNN	General		√		√		√ (synthetic)		
Lahoud et al. [48]	2019	CNN	General			√		√		√ (transfer)	
FusionGAN [32]	2019	**GAN**	VIF	√	√		√			√	√ (Tensorflow)
DDcGAN [49]	2019	GAN	VIF	√	√		√			√	
MD-WGAN [50]	2019	GAN	VIF	√	√		√			√	
FusionNet [22]	2019	GAN	General	√	√		√		√ (fused images)		
NestFuse [51]	2020	AE	VIF	√		√	√			√	√ (Pytorch)
Zhu et al. [52]	2020	AE	VIF			√	√			√	
Patel et al. [53]	2020	AE	VIF			√	√			√	
DIDFuse [54]	2020	AE	VIF		√		√			√	√ (Pytorch)
PFAF-Net [55]	2020	AE	General			√	√			√	
Feng et al. [56]	2020	CNN	VIF			√	√		√ (synthetic)		
Li et al. [57]	2020	CNN	VIF			√	√			√ (transfer)	
An et al. [58]	2020	CNN	VIF		√		√		√ (fused images)		
VIF-Net [59]	2020	CNN	VIF		√		√			√	
Mustafa et al. [60]	2020	CNN	VIF		√		√			√	
IFCNN [23]	2020	CNN	General		√		√		√ (synthetic)		√ (Pytorch)
DIF-Net [25]	2020	CNN	General		√		√			√	
FusionDN [61]	2020	CNN	General		√		√			√	√ (Tensorflow)
PMGI [62]	2020	CNN	General		√		√			√	√ (Tensorflow)
U2fusion [63]	2020	CNN	General		√		√			√	√ (Tensorflow)
Zhao et al. [64]	2020	CNN	General		√		√			√	
Xu et al. [65]	2020	GAN	VIF	√	√		√			√	
AttentionFGAN [66]	2020	GAN	VIF		√		√			√	
D2WGAN [67]	2020	GAN	VIF		√		√			√	
LBP-BEGAN [68]	2020	GAN	VIF		√		√			√	
FLGC-FusionGAN [69]	2020	GAN	VIF		√		√			√	
Zhao et al. [70]	2020	GAN	VIF	√	√		√			√	
Detail-GAN [6]	2020	GAN	VIF	√	√		√			√	√ (Pytorch)
DDcGAN [71]	2020	GAN	General		√		√			√	√ (Tensorflow)
Fu et al. [72]	2021	AE	VIF			√	√			√	
SEDRFuse [28]	2021	AE	VIF			√	√			√	√ (Tensorflow)
RFN-Nest [73]	2021	AE	VIF			√	√			√	√ (Pytorch)
Fu et al. [74]	2021	AE	VIF			√	√			√	
CSF [75]	2021	AE	VIF			√	√			√	√ (Tensorflow)
AUIF [76]	2021	AE	VIF				√			√	
SFA-Fuse [77]	2021	AE	VIF	√		√	√			√	
DepthFuseNet [78]	2021	AE	VIF			√	√			√	
DenseNetFuse [79]	2021	AE	VIF			√	√			√	
UNFusion [80]	2021	AE	VIF			√	√			√	√ (Pytorch)
SMoA [81]	2021	AE	VIF			√	√			√	√ (Pytorch)
IFSepR [82]	2021	AE	General			√	√		√ (synthetic)		
RXDNFuse [83]	2021	CNN	VIF	√	√		√			√	
VMDM-fusion [84]	2021	CNN	VIF			√	√			√ (transfer)	
Ren et al. [85]	2021	CNN	VIF			√	√			√	
STDFusionNet [31]	2021	CNN	VIF		√		√		√ (mask)		√(Tensorflow)
IR-MSDNet [86]	2021	CNN	VIF		√		√			√	
DRF [87]	2021	CNN	VIF		√		√			√	√ (Tensorflow)
Xu et al. [88]	2021	CNN	VIF		√		√			√	
TSFNet [89]	2021	CNN	VIF		√		√			√	
MLFusion [27]	2021	CNN	VIF		√		√			√	√ (Pytorch)
Jian et al. [90]	2021	CNN	VIF		√		√			√	
HAF [91]	2021	CNN	General		√		√			√	
Liu et a . [92]	2021	CNN	General			√	√			√	
SDNet [93]	2021	CNN	General		√		√			√	√ (Tensorflow)
FusionADA [94]	2021	GAN	VIF		√		√		√ (mask)		
SSGAN [95]	2021	GAN	VIF		√		√		√ (mask)		
RCGAN [24]	2021	GAN	VIF	√	√		√		√ (fused images)		√ (Pytorch)
SDDGAN [96]	2021	GAN	VIF	√	√		√		√ (mask)		√ (Tensorflow)
TC-GAN [33]	2021	GAN	VIF	√		√	√			√	
Bhagat et al. [97]	2021	GAN	VIF		√		√			√	
MgAN-Fuse [98]	2021	GAN	VIF		√		√			√	
MFEIF [99]	2021	GAN	VIF		√		√			√	√ (Pytorch)
Perception-GAN [100]	2021	GAN	VIF		√		√			√	
GANMcC [101]	2021	GAN	VIF		√		√			√	√ (Tensorflow)
Laplacian GAN [26]	2021	GAN	VIF		√		√			√	
LatRAIVF [102]	2021	GAN	VIF		√		√		√ (synthetic)		
GAN-FM [103]	2021	GAN	VIF	√	√		√			√	√ (Tensorflow)
DNDT [35]	2021	**Transformer**	VIF	√		√	√			√	
MAFusion [104]	2022	AE	VIF			√	√			√	
ERNet [105]	2022	AE	VIF	√		√	√			√	
Res2Fusion [106]	2022	AE	VIF		√		√			√	√ (Pytorch)
CUFD [107]	2022	CNN	VIF		√		√			√	
SeAFusion [108]	2022	CNN	VIF		√		√		√ (application)		√ (Pytorch)
CLF-Net [109]	2022	CNN	VIF		√			√		√	√ (Pytorch)
PIAFusion [110]	2022	CNN	VIF		√		√			√	√ (Pytorch)
UMF-CMGR [111]	2022	CNN	VIF		√		√			√	√ (Pytorch)
IPLF [112]	2022	CNN	VIF	√	√		√		√ (synthetic)		√ (Pytorch)
StyleFuse [113]	2022	CNN	VIF	√	√		√			√	
MHTNet [114]	2022	CNN	VIF		√		√			√	
TarDAL [115]	2022	GAN	VIF	√	√		√		√ (application)		√ (Pytorch)
DCDR-GAN [116]	2022	GAN	VIF			√		√		√	
CGTF [117]	2022	Transformer	VIF		√		√			√	
YDTR [118]	2022	Transformer	VIF		√		√			√	√ (Pytorch)
SwinFuse [119]	2022	Transformer	VIF			√		√		√	√ (Pytorch)
MFST [120]	2022	Transformer	VIF	√		√	√			√	
SwinFusion [121]	2022	Transformer	General		√		√			√	√ (Pytorch)



(a) Unsupervised methods.



(b) Supervised methods.

Fig. 6. Main architectures of CNN-based VIF methods. Note that the CNN-based model can contain single branch or dual branches. It can also contain other components, e.g., image decomposition.

(1) CNN is applied to a part of the VIF process

In many unsupervised CNN-based methods, CNN is only applied to a part of the VIF process. For example, Liu et al. [46] decomposed source images into a base part and a detail part. Then, they fused the base parts using a weighted summation method and fused the detail parts using CNNs and a multi-layer feature fusion strategy. Hou et al. [59] proposed VIF-Net that uses CNN in feature extraction and image reconstruction.

(2) CNN is applied to the whole VIF process

In some unsupervised CNN-based methods, CNN is applied to the entire VIF process. For example, Xu et al. [107] and Mustafa et al. [60] applied CNN in all three stages of VIF.

(3) Measures to improve performance

Various measures have been taken to improve the performance of CNN-based VIF methods, as summarized below:

Residual Connection. Residual connection [127] is a crucial and effective technique in deep learning. It was introduced to the VIF field by Li et al. [43] in 2019. Afterwards, a lot of VIF methods [27], [60], [83], [87], [128] utilized residual connection and exhibited good performance. For example, Long et al. [83] proposed a VIF method based on a dense residual network that combines ResNet with DenseNet. Li et al. [114] used residual connection in both encoder and decoder. Mustafa et al. [60] proposed a residual self-attention block to refine features.

Dense Connection. Dense connection [129] has shown good performance in many applications. It was introduced to the VIF field by Li et al. [42] in 2019, where dense connections were used in the encoder to improve representation ability. Since then, dense connections have been widely applied to VIF methods [57], [59], [80], [83], [85], [86], [89], [107], [130] to improve performance. Most methods applied dense connections only in the feature extraction stage [42], [57], [83], [85], [130], while some methods [80] applied dense connections in both feature extraction and image reconstruction stages.

Attention Mechanism. Incorporating an attention mechanism is one of the most frequently used techniques to improve performance in VIF. Different types of attention mechanisms have been utilized, including channel attention [45], [52], [89], [128] and spatial attention [80], [128]. Attention mechanism can be applied to different stages of VIF methods. In some studies, attention mechanism is used to refine features [52], [57], [60], [128]. For example, Zhu et al. [52] utilized a multi-channel attention mechanism to improve features, Li et al. [57] extracted features from source images using a pre-trained CNN and then adopted the attention mechanism to enable more efficient feature extraction, Mustafa et al. [60] utilized a self-attention mechanism to refine visible and infrared features. In some other studies, attention mechanism is used as the fusion strategy [131]. Moreover, attention mechanism is used to perform feature refinement and feature fusion in some methods. For example, Wang et al. [80] utilized a L_p normalized attention-based strategy to refine and combine deep features.

Multiscale Features. Multiscale features have been frequently used in CNN-based VIF methods [52], [80], [132]. One way of achieving multiscale features is using convolution kernels with different sizes because smaller kernels can effectively extract low-frequency information, while larger kernels can capture large features [52]. For example, Wang et al. [80] and Liu et al. [132] used two different kernel sizes, i.e., 1×1 and 3×3 kernels, to extract multiscale features.

Multilevel Features. One way of obtaining multilevel features is by using image decomposition methods. For example, Yan et al. [133] first decomposed source images into a base layer and a detail layer. The base layers were fused using visual saliency weight map method, while the detail layers were first processed by multi-resolution singular value decomposition to generate multi-level features and then fused by a pre-trained VGG19. Another way of implementation is by using features from different layers of a CNN, as was first done by Li et al. [30]. Since then, fusing multilevel features has become a commonly-used technique to improve fusion performance in many VIF methods [80], [107], [114].

Contrastive Learning. In VIF, salient objects in the infrared image and background details in the visible image are often kept in the fused image [31]. Therefore, the fused image is usually similar to the infrared image in salient objects and the visible image in background details. Based on this idea, contrastive learning can be applied to VIF. For example, Zhu et al. [109] applied contrastive learning to VIF by designing a contrastive learning framework and a contrastive loss. In addition, Luo et al. [82] proposed a contrastive difference loss to perform feature disentanglement so that the distinction between the common and private features of source images is maximized. This method can be applied to several image fusion tasks.

Neural Architecture Search. Most researchers focus on designing different architectures to perform VIF. This is heavily dependent on the experience of researchers and may take a lot of time. It is therefore desirable to learn architectures automatically. To this end, Liu et al. [81] proposed SMoA, which utilizes Neural Architecture Search (NAS) to discover modality-oriented feature representation for visible and infrared

images. This method can alleviate the issue of designing architectures manually. To the best of our knowledge, this is the first method that aims to search VIF network architectures automatically. The same team did further research in this direction. For example, they proposed an architecture search scheme in another VIF method [91]. Specifically, they first constructed a hierarchically aggregated fusion architecture and then established a search space in which latent architectures can be searched. Moreover, they proposed a lightweight architecture based on NAS [134]. In our opinion, this is a promising direction in the future because it can save human efforts of designing architectures manually and has the potential to find better architectures.

Image or Feature Decomposition. Some methods directly fuse the whole source images, while some methods first decompose source images into different parts and then fuse these different parts using different strategies. For example, Liu et al. [46] decomposed source images into base part and detail part. They then fused the base parts using a weighted summation method, while fused the detail parts using CNNs and a multilayer feature fusion strategy. Apart from image decomposition, some methods perform feature decomposition. For example, Xu et al. [87] proposed to learn disentangled representation for visible and infrared images. Specifically, scene-based and attribute-related representations are learned for both visible and infrared images. Then, different fusion strategies are applied to fuse scene-based and attribute-related features, respectively. The fused image is obtained via a CNN based on residual block and deconvolution layers. Different from [87], Xu et al. [107] performed feature map decomposition to obtain common and unique parts. Then, they applied different fusion rules to fuse the common parts and the unique parts.

Illumination-Aware Module. Illumination conditions significantly affect the reliability of visible and infrared images. Therefore, it is important to consider illumination conditions in the image fusion process. Tang et al. [110] proposed an illumination-aware module to evaluate the illumination conditions, which were then used to guide the image fusion process via an illumination-aware loss.

Other Types of Convolutions. Instead of regular convolutions, some studies proposed to use other types of convolutions in the VIF methods because those convolutions have special properties. For example, Mustafa et al. [60] utilized dilated convolution to increase receptive fields.

The measures mentioned above are important building blocks in deep learning-based VIF methods. Some of these measures have been taken together in many VIF methods to improve performance. For example, Li et al. [57] proposed a VIF method based on DenseNet and attention, Mustafa et al. [60] utilized multiscale features and residual connection to improve performance, Long et al. [83] proposed a VIF method based on a dense residual network that combines ResNet and DenseNet, Zou et al. [135] employed attention and multiscale features in their method, Shen et al. [136] used attention, residual connection, and densenet in their method.

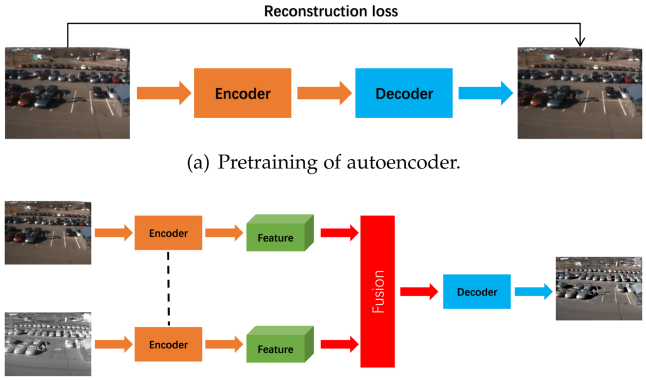
2) Supervised Methods: Although most CNN-based methods are unsupervised, there are a few supervised methods, whose overall architecture is shown in Fig. 6(b). Several types of 'ground truth' are used in supervised CNN-based methods.

The first type is fused images generated by other methods. For example, An et al. [58] proposed a CNN-based VIF method consisting of a coding layer, fusion layer, decoding layer, and output layer. They generated training labels using the method proposed by Zhang et al. [4]. The second type is the synthetic ground truth of other image fusion tasks. For example, Liu et al. [29] trained their model using high-quality images and their blurred versions generated using multiscale Gaussian filtering and random sampling. In addition, Feng et al. [56] proposed a VIF method whose fully convolution network was also trained using clear RGB images and their blurred versions. Moreover, Wang et al. [47] used clear RGB images and their blurred versions to finetune pre-trained VGG and ResNet models. Zhu et al. [112] generated synthetic training data using both visible and infrared images. The third type is object mask. In VIF, it is very important to highlight targets that are more visible in infrared images. Therefore, some methods try to improve fusion performance with the help of object mask. For example, Ma et al. [31] defined the VIF process as the fusion of texture information in the visible image and salient target in the infrared image. As a result, they annotated target masks in source images and used these masks to help the network for salient target detection and information fusion. The fourth type, which was used recently, is the ground truth of downstream applications. For example, Tang et al. [108] used the ground truth of scene segmentation in the training. This type will be further discussed in Section IV-H.

In summary, supervised methods can have similar network architectures as unsupervised methods. However, because there is no real ground-truth fused image in the VIF task, different kinds of 'ground truth' should be used to compensate for this and to compute the loss function. Moreover, the measures used to improve the performance of unsupervised CNN-based VIF methods (see Section II-D1) are also frequently used in supervised methods.

3) Transfer Learning-Based Methods: A part of CNN-based methods is based on transfer learning, where off-the-shell models pre-trained (usually by other researchers) on large-scale datasets, such as ImageNet, are used to extract features. For example, Li et al. [43] and Zhang et al. [137] used ResNet50 to extract features from the high-frequency parts of source images. Li et al. [30] and Ren et al. [41] used VGG19 to extract features. Yang et al. [84] employed VGG16 and Li et al. [57] used DenseNet-201 pre-trained on ImageNet to extract deep features. In some studies, these features are processed before fusion. For example, Li et al. [43] applied zero-phase component analysis and l_1-norm to normalize the extracted features.

In these transfer learning-based methods, fusion rules are usually designed manually to fuse the extracted features. Regarding image reconstruction, some studies first generate a weight map [43] or multiple weight maps [57], [84] based on extracted features, and then use a weighted average strategy to obtain the fused image. Alternatively, some studies [30], [137] combine the fused base part and detail part to obtain the fused image. In



(a) Pretraining of autoencoder.

(b) Autoencoder-based VIF methods.

Fig. 7. Basic architectures of AE-based VIF methods. First, an autoencoder is trained to reconstruct the input image. Then, the trained encoder and decoder are used to generate a fused image with the help of a fusion rule.

addition, Ren et al. [41] utilized an L-BFGS method to optimize a loss function computed based on the extracted features.

E. Autoencoder-Based VIF Methods

Autoencoder (AE)-based methods consist of two steps. In the first step, an autoencoder is pre-trained using visible images and/or infrared images, as shown in Fig. 7(a). In the second step, the trained encoder is used for feature extraction, and the trained decoder is used for image reconstruction, as shown in Fig. 7(b). The fusion between encoder and decoder is usually performed according to manual fusion rules or learned through a second training step using visible-infrared image pairs. Note that the difference between AE-based methods and transfer learning-based methods (Section II-D3) is that an autoencoder is trained in AE-based methods, while off-the-shelf pre-trained models are directly used in transfer learning-based methods.

A well-known and pioneering AE-based VIF method is DenseFuse [42], which uses MS-COCO [138] to pretrain the AE and different fusion strategies (addition and l_1-norm) to perform feature fusion. Raza et al. [55] also proposed an AE-based VIF method that uses weight map-based fusion strategy. The weight map is obtained using l_1-norm and softmax based on extract features. Moreover, Fu et al. [72] proposed an AE-based method that has dual branches in the encoder. An addition strategy and a channel selection strategy are utilized for feature fusion. Furthermore, Jian et al. [28] proposed SEDRFuse, which first trains an AE using both visible and infrared images to obtain an encoder that can extract features and a decoder that can reconstruct fused image. They designed an attention-based feature fusion strategy to fuse intermediate features and a choose-max strategy to fuse compensation features. Similarly, Li et al. [51] proposed to use a fusion strategy based on spatial attention and channel attention to fuse multiscale features. Wang et al. [106] also proposed a fusion strategy based on spatial non-local attention and channel no-local attention to handle long-range information. Zhao et al. [76] proposed a slightly different AE-based method, using a two-scale decomposition process to decompose source images into base part and detail part. Addition strategy is utilized to fuse the decomposed feature maps of the base parts (and the detail parts) from both source images.

However, the feature fusion step of the above-mentioned AE-based methods is performed according to manual fusion rules, which may not be very effective. To solve this issue, Li et al. [73] recently proposed RFN-Nest that utilizes a residual fusion network learned via training using visible and infrared image pairs to perform feature fusion. In addition, they utilized multiscale features in the encoder and nest connection in the decoder to improve performance. Similarly, Xu et al. [75] first used both visible and infrared images to train an autoencoder. Then, they trained a classifier to obtain a classification saliency map for features, which was then used to fuse features in a pixel-level weighting manner. A decoder was finally employed to reconstruct the fused image based on fused features.

Many AE-based methods [42], [51], [52], [55], [74], [106] only utilize RGB images to train the AE and directly apply the trained AE to infrared images. Therefore, the performance may be limited due to the difference between RGB and infrared images. To solve this issue, some methods [28], [53], [72], [75], [76], [78] use both RGB and infrared images to train the AE in turn, while some methods [105] use RGB-infrared image pairs to train the encoder. In addition, some methods [79] train a visible AE and an infrared AE. Moreover, in SFA-Fuse [77], an encoder and two decoders are utilized in the pretraining stage to alleviate the vital information loss. Specifically, the two decoders are used to reconstruct the visible image and the infrared image, respectively. Furthermore, Liu et al. [81] proposed to use two encoders and one unified decoder, whose architectures are obtained using NAS techniques. Another way is using visible-infrared image pairs to train a fusion module [73], [75]. Finally, it is worth mentioning that the measures used to improve performance introduced in Section II-D can also be applied to AE-based methods, such as multiscale features [52], [80], multi-level features [139], dense connections [79], [80], [140], attention mechanism [80], and other types of convolutions (depthwise convolution) [53], [78].

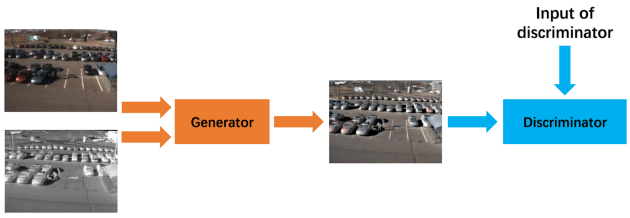
F. GAN-Based VIF Methods

In 2019, Ma et al. [32] introduced GAN to image fusion. Since then, many GAN-based VIF methods have been proposed, and GAN-based methods have become one of the most important types of VIF methods. We will introduce representative GAN-based VIF methods in this section.

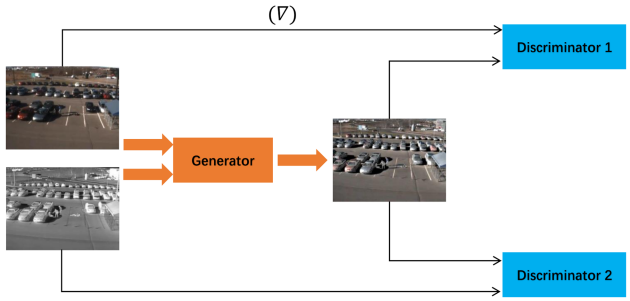
1) Unsupervised Methods: The majority of GAN-based VIF methods are unsupervised methods. The training is usually driven by a loss function that compares the difference of the fused image with the source images. This loss function usually contains several items that reflect the difference from different perspectives. In this section, we discuss these methods based on the number of generators and discriminators, as shown in Fig. 8.

(1) One generator and one discriminator

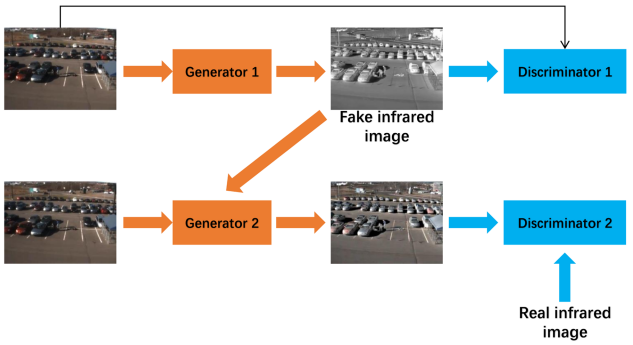
FusionGAN [32] is the first GAN-based VIF method. In FusionGAN, a generator is designed to generate the fused image with highlighted targets, and a discriminator is designed to force



(a) One generator and one discriminator [6], [33], [66], [69], [70]. The generator is used to generate fused images, and the discriminator is used to make the fused images similar to either the visible or the infrared image.



(b) One generator and two discriminators [50], [51], [67], [68], [72], [99]. The two discriminators are used to make the fused image contain features from both visible and infrared images.



(c) Two generators and two discriminators [71]. Generator 1 is used to generate a fake infrared image, and Generator 2 is used to produce fused image based on the visible image and the fake infrared image.

Fig. 8. Basic architectures of unsupervised GAN-based VIF methods.

the fused image to contain more textural details as in visible images. Ma et al. [6] extended FusionGAN by introducing a detail loss and a target edge-enhancement loss to let the fused images have more texture details. FusionGAN was also improved by Yuan et al. [69] using learnable group convolution to replace convolutional layers in the generator to reduce computational costs and adopting multilevel residual network containing dense blocks in the generator to enhance network capacity.

In GAN-based VIF methods, some measures have also been used to improve fusion performance. For example, Xu et al. [68] utilized a local binary pattern loss for training. Xu et al. [65] utilized residual blocks and skip connections in the generator. Fu et al. [100] proposed to utilize dense blocks to allow the generator to learn more information. They not only concatenate the features of shallow layers and deeper layers but also insert the visible image at each layer of the generator to help the network learn visible information. Furthermore, other measures, such as attention mechanisms [141] and residual connections [65], have also been applied.

The above-mentioned methods directly use original visible and infrared images as the input to the generator. In addition, these methods only consider the main information, i.e., textures in visible images and contrast information in infrared images, but ignore the auxiliary information, i.e., textures in infrared images and contrast information in visible images. To solve this issue, Ma et al. [101] proposed GANMcC that uses a specific content loss for the generator. They use a two-branch architecture for the generator, and each branch (a gradient branch and a contrast branch) adopts a different combination of source images as the input. The input to the gradient branch is a concatenation of two visible images and one infrared image, while the input to the contrast branch is a concatenation of two infrared images and one visible image. This design enables the generator to get main and auxiliary information from both source images.

In addition to the above methods, an interesting VIF method is the MFEIF proposed in [99]. One interesting feature of MFEIF is that it does not need well-aligned image pairs to train. This method also utilizes multiscale features through a coarse-to-fine deep architecture. In addition, a cross-domain edge-guided attention mechanism is designed to encourage the model to focus on common structures, thus keeping more details. Another interesting method is proposed by Liao et al. [142] who employed VGG19 to extract features from the visible image and the fused image generated by the generator, and then minimized the Wasserstein distance in the feature space.

The main shortcoming of most GAN-based methods mentioned above is that only one discriminator is adapted to force the generated fused image to be similar to either the visible image [6], [32], [65], [69], [97], [100] or infrared image [101]. However, in either way, the fused image will lose some details of source images as the adversarial game proceeds. To solve this issue, Ma et al. [101] proposed to use a multiclassification-based discriminator to achieve a balance between the visible distribution and infrared distribution. Moreover, some researchers proposed to use more discriminators to solve this issue.

(2) One generator and more discriminators

To solve the issue of considering a single source image in the discriminator, some researchers extended the GAN-based methods to two or more discriminators. The main advantage of using more discriminators is to preserve features in both source images. For example, Xu et al. [49] proposed the DDcGAN, which is a GAN-based VIF method with two discriminators that can be employed to preserve features in both source images. Another novelty of this method is that it can fuse source images of different resolutions. This method is then extended by Ma et al. [71]. From technical point of view, the main improvements are as follows. First, a densely connected CNN is used to replace the U-Net in the generator. Second, the discriminator takes the image itself rather than the gradients of images as input. Third, a deconvolution layer is used to replace two upsampling layers

to perform the upsampling operation on the infrared image to generate the input for the generator.

Other researchers also noticed that adopting two discriminators is beneficial. For example, Li et al. [50], [66], [67], [98] designed a series of GAN-based VIF methods using one generator and two discriminators. Initially, they proposed MD-WGAN [50] that uses one generator and two discriminators. A texture loss function based on local binary patterns was defined to force the fused image to keep more texture information. This method was then extended to D2WGAN [67]. The main improvement is that a GAN with the Wasserstein distance is adopted. This method was then extended to MgAN-Fuse [98] that employs multiscale attention in the encoder. Also, the source visible image and infrared image are processed using two different encoders instead of being concatenated and processed by the same CNN. Li et al. [66] also proposed AttentionFGAN, which employs multiscale attention mechanism in both generator and discriminators. In addition, in the generator, an intention map of visible image and an intention map of infrared image are generated by two multiscale attention networks, respectively. These two intention maps and source images are used together to generate the fused image. This is different from MgAN-Fuse [98] that adopts an encoder-decoder architecture in the generator. Furtheremore, an attention loss was designed between the fused image and source images based on discriminators. Moreover, Zhang et al. [103] proposed to use one full-scale skip connection-based generator and two Markovian discriminators to keep useful information in visible and infrared source images.

In addition to the methods using two discriminators, Song et al. [143] recently used one generator and three discriminators in their VIF method. Apart from the visible and infrared discriminator, they designed a difference image discriminator to account for the difference between visible and infrared images.

(3) Two generators and two discriminators

Zhao et al. [70] proposed a VIF method using two generators and two discriminators. They first generate a fake infrared image from the visible image using the first generator. Then, they fuse the fake infrared image and the visible image to obtain the fused image using the second generator. The first discriminator is used to compare the fused image with the visible image. The second discriminator is used to compare the fused image, the real infrared image, and the fake infrared image. To the best of our knowledge, this is the first VIF method using only visible images as input in the test stage.

2) Supervised Methods: There are a few supervised GAN-based methods that use different types of ground truth. The first type uses fused images generated by other methods as ground truth. For example, Lebedev et al. [22] proposed a method consisting of one generator and one discriminator. They used fused images generated using Laplacian pyramid algorithm accompanied by MultiScale Retinex [144] as ground truth. Afterwards, Li et al. [24] proposed RCGAN based on coupled GAN [145]. In particular, RCGAN has two generators and two discriminators. An innovation of this method is that pre-fused images generated by GFF [146] are adopted to be optimized in the coupled generators. However, the performance of RCGAN will be affected by the chosen method that generates the pre-fused images.

The second type uses object masks as supervised signals. For example, Gu et al. [94] proposed FusionADA consisting of one generator and one discriminator. In FusionADA, labeled masks are employed to help the fused image contain salient thermal targets from the infrared image. However, the input to the discriminator only contains salient thermal targets, thus the fused image may loss texture details of the visible image. Afterwards, Hou et al. [95] proposed SSGAN that utilizes semantic segmentation to generate target masks. A dual-encoder-single-decoder structure containing a foreground path and a background path is employed in the generator to generate fused images. The input to the discriminator is a combination of visible image background and infrared image foreground. In this way, thermal targets in infrared images and texture details in visible images can be preserved. However, in SSGAN, the segmentation network needs to be trained separately. Recently, Zhou et al. [96] proposed a semantic-supervised VIF method using one generator and two discriminators. A novel contribution of that work is that an information quantity discriminator block was designed to generate fusion weights, which are then used to preserve semantic information in both visible and infrared images. Moreover, the two discriminators help to make the fused images contain the texture details of the visible image and thermal radiation of the infrared image. Labeled masks are required to train the model.

The third type uses the Y channel in the YCbCr space of RGB images from an RGB-D dataset as ground truth [102]. This method generates synthetic infrared and visible images based on the ground-truth image [102]. The synthetic dataset is then used for training.

In summary, GAN-based methods have become one of the most popular types of VIF methods. Note that GAN is applied to only a part of the image fusion process in some GAN-based methods. For example, GAN is only used to fuse detail layers of source images in Laplacian GAN [26].

G. Transformer-Based VIF Methods

Transformers can handle long-range dependencies and have been applied to various natural language processing [147] and vision tasks [148], [149]. In 2021, transformers were introduced to the image fusion field, and some transformer-based methods have been proposed to perform VIF [35], [37], [117], [118], [119], [120], [150], [151], general image fusion [34], [36], [121], [152], and other image fusion tasks [153], [154], [155], [156].

In some methods, transformer is only applied to perform feature fusion. For example, VS et al. [34] proposed to use a transformer-based multiscale fusion strategy to fuse both local and global information. Zhao et al. [35] proposed DNDT, which uses a dual transformer as the fusion strategy. An encoder was designed to extract features from source images and a decoder was designed to construct the fused image. Liu et al. [120]

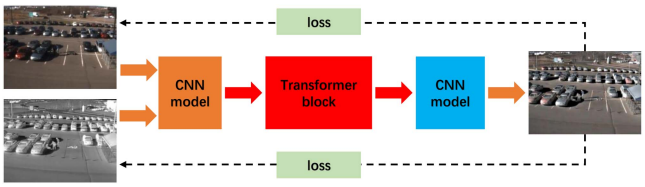


Fig. 9. An example architecture of transformer-based VIF methods. A CNN model is used to extract features, which are fused using a transformer block. The fused image is obtained by using another CNN model.

proposed a transformer fusion block based on focal self-attention to fuse multiscale features extracted by a multiscale encoder. The fused image is obtained through a decoder containing nest connections. Recently, Rao et al. [37] proposed a VIF method based on transformer and GAN. Specifically, a spatial transformer and a channel transformer are combined to form a transformer fusion module in the generator. Moreover, Ma et al. proposed SwinFusion [121], which is a general image fusion method based on swin transformer. They showed that global information is crucial in image fusion and visualized global information. To the best of our knowledge, SwinFusion is the first study that clearly demonstrates global information in the context of image fusion.

In some other methods, transformer is also applied to other stages of VIF methods. For example, Fu et al. [36] proposed a pyramid patch transformer method consisting of a transformer-based feature extraction module and an MLP-based decoder. This method is designed as an AE-based method. The average strategy is used for fusion. Similarly, Wang et al. [119] designed SwinFuse in an AE-based framework and used a transformer-based encoder to extract global features. Tang et al. [118] proposed YDTR that combines CNN and transformer in both encoding branches and decoding branch. Features from visible images and infrared images are added together. A similar method is the CGTF [117] that uses transformer feature extraction module and convolution feature extraction module in both encoding branches and decoding branch. In addition, Tang et al. [151] designed a transformer-based global feature extraction branch, which is parallel to their CNN-based local feature extraction branch. Yang et al. [150] used a stack of transformer blocks and convolution blocks to generate fused image from source images.

The architecture of transformer-based methods varies from one method to another. It is worth mentioning that very few existing transformer-based methods are purely based on transformer. CNNs are usually used together with transformers in the image fusion process, as shown in the example architecture given in Fig. 9. Moreover, all existing transformer-based VIF methods are unsupervised methods. Therefore, the loss function is computed using the fused image and source images.

H. Other Deep Learning-Based VIF Methods

In addition to CNN-based methods, AE-based methods, GAN-based methods, and transformer-based methods, there are also some other deep learning-based VIF methods. For example, Wu et al. [21] proposed a DBM-based VIF method in 2018. However, the number of other deep learning-based VIF methods is very limited.

III. General Image Fusion Methods

The majority of the above-mentioned image fusion methods are designed specifically for the VIF task. In addition to those methods, researchers have also designed general image fusion methods that can be applied to several image fusion tasks, i.e., VIF, MFIF, MEF, and MEDIF. In general image fusion methods, the same model is usually employed to perform different fusion tasks. General image fusion methods also contain CNN-based methods [23], [61], [62], [63], AE-based methods [82], GAN-based methods [71], [157], [158] and transformer-based methods [34], [36], [121].

There are different ways to implement general image fusion methods. First, some methods [62], [136] choose different weights for loss terms in different image fusion tasks. For example, Zhang et al. [93] utilized a squeeze-and-decomposition network and manually adjusted weights in the intensity loss to satisfy the requirements on intensity distribution in different image fusion tasks. Second, different feature fusion strategies are applied to different image fusion tasks. An example is the IFCNN proposed by Zhang et al. [23], which uses elementwise-mean strategy for MEF and elementwise-max strategy for VIF, MFIF, and MEDIF. Third, some researchers [48] use a pre-trained model on other computer vision tasks, such as image classification, to extract features for various image fusion tasks. Furthermore, apart from these studies, other general image fusion methods have also been proposed [25], [64], [71]. Note that some general image fusion methods use training data for one task and then apply the trained model to various image fusion tasks [23], [82]. In contrast, some methods use different training data for each image fusion task [63], [93], [121].

General image fusion methods are convenient to use because they can perform several image fusion tasks. Some methods can also utilize the common characteristics between various image fusion tasks. However, different image fusion tasks have very different characteristics and thus have different key points to consider to obtain good fusion performance. For example, in VIF, it is crucial to retain texture details in visible images and salient information in infrared images. In MFIF, it is essential to find the boundary between focused and defocused region and handle the defocus spread effect (DSE) properly. In MEF, it is crucial to remove halo effect and ghost effect. Indeed, as claimed by Zhang [123], it is challenging for general image fusion methods to handle these key differences between different image fusion tasks in a single model. Therefore, general image fusion methods may not have superior performance over specific VIF methods. More details about general image fusion methods can be found in [123], [124].

IV. Recent Development Characteristics

This section summarizes some new characteristics of the VIF field shown in recent years.

A. More and More Types of Deep Learning Models Have Been Applied to VIF

When deep learning was introduced to VIF, only DBM [21] and CNN [29], [30], [41] were applied to perform VIF. At a later time (2019), GANs [32] and AE [42] were introduced to VIF and became very important types of VIF methods. In 2021, variational autoencoder [85] and transformer [35], [36], [37] were also introduced to this field. Moreover, some important network architectures, such as DenseNet and ResNet, have been introduced to VIF and become important building blocks. We believe that more and more types of deep learning models will be introduced to perform VIF for better fusion performance.

B. Most Methods are Unsupervised Methods

Most deep learning-based VIF methods are unsupervised methods because there are no ground truth images in VIF. This is different from MFIF and MEF, where all-clear and well-exposed images can be used as ground truth for synthetic datasets. Therefore, in VIF, much attention of researchers has been paid to design various loss functions based on the fused image and source images.

C. Combination of Deep Learning and Traditional Image Processing Techniques

It is not uncommon that deep learning and traditional image processing techniques are utilized together in a VIF method. For example, Raza et al. [86] proposed to use a dense multiscale network and quadtree decomposition as well as Bezier interpolation to extract different features. Some methods also combine GAN with traditional image processing techniques. For example, in the method of Wang et al. [26], source images were first decomposed into base layers and detail layers. Then, the base layers were fused using a Laplacian pyramid method, while the detail layers were fused using a GAN. In addition, Yang et al. [33] proposed the TC-GAN that combines GAN and an adaptive guided filter (AGF). Specifically, the generator is designed to generate a combined texture map, which is adopted as a guidance image of the AGF. By combining deep learning and traditional image processing techniques, the advantages of them can be kept in the VIF method.

D. Combination of VIF with Other Tasks

Previously, VIF was the sole objective in almost all VIF studies. In recent years, some researchers have investigated performing VIF with other tasks together. For example, Gu et al. [159] proposed a dataset-free self-supervised image super-resolution fusion method that can perform VIF and super-resolution in a single network. Specifically, two low-resolution source images can be fused to a high-resolution image. Li et al. [27] also proposed a VIF method that can perform VIF and super-resolution together. Xiao et al. [160] proposed a knowledge distillation method for simultaneous VIF and super-resolution. By combining VIF and other tasks, the model can be used more efficiently because one model can perform more tasks.

E. Learning Image Fusion and Registration Together

Because of the different imaging mechanisms of visible and infrared images and the different parameters of visible and infrared cameras, it is difficult to accurately align visible-infrared image pairs. Many methods have been proposed to perform visible-infrared image registration [161], [162], [163]. However, almost all of these studies do not consider the image fusion task. To solve this issue, some researchers started to learn image fusion and registration together [111], [164]. For example, Wang et al. [111] first generates a pseudo-infrared image using a cross-modality perceptual style transfer network. Then, they learn the displacement vector field between the real infrared image and the pseudo-infrared image, which is an easier mono-modality registration problem. The learned displacement vector field is then used to reconstructed registered real infrared image. Finally, they perform image fusion using visible image and the registered infrared image via a dual-path interaction fusion network. A loss function consists of style transfer loss, cross regularization loss, registration loss and image fusion loss is designed to guide model training. Moreover, Xu et al. [164] proposed RFNet, which is a mutually reinforcing framework that learns fusion and registration together. Specifically, RFNet utilizes image fusion to provide feedback for image registration. However, although the idea is inspiring, RFNet is designed for the registration of visible images and near-infrared images instead of thermal infrared images.

F. VIF Methods for Images of Different Resolutions

Most existing VIF methods aim to fuse visible and infrared images of the same resolution. However, it is more often to have high-resolution visible images and low-resolution infrared images in practice. Recently, some researchers started developing methods to fuse images of different resolutions. For example, Xu et al. [49] proposed a method to fuse a high-resolution visible image and a low-resolution infrared image. This is achieved by first upsampling the infrared image with two upsampling layers and then adopting a discriminator to differentiate the original infrared image and a downsampled fused image. This method was then extended by Ma et al. [71]. Specifically, a deconvolution layer, which learns a mapping from low-resolution feature to high-resolution feature, is used to replace the two upsampling layers. In this way, the parameters were obtained from training instead of pre-defining. However, these two methods have a requirement on the input size of source images, i.e., the ratio between the resolution of the visible image and the infrared image should be 4.

Recently, Li et al. [27] proposed a meta-learning-based VIF method that can fuse source images of different resolutions to a fused image with arbitrary resolution. Liu et al. [165] also proposed to fuse source images of different resolutions. They designed up-projection and down-projection blocks to achieve feature mapping between images of different resolutions. This method has been used to fuse visible and infrared images and multi-resolution medical images.

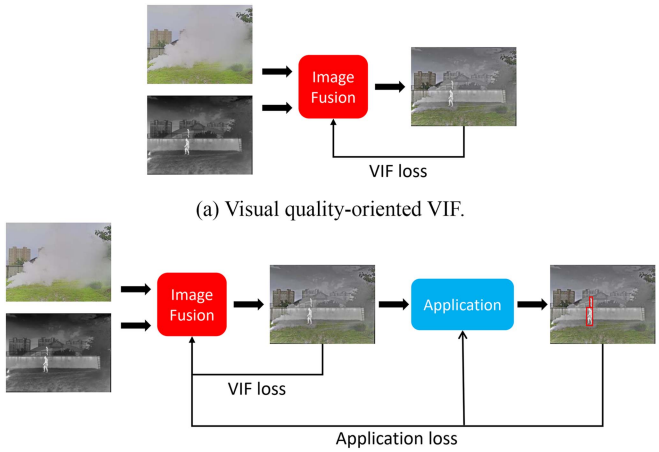


Fig. 10. Visual quality-oriented VIF v.s. application-oriented VIF. Person detection is used as an example. The visible and infrared images are provided in the M^3FD dataset [115]. The fused image is generated by the authors using the MGFF [18] image fusion algorithm.

G. Research on Benchmarks

Unlike many tasks in computer vision, image fusion has suffered from the lack of benchmarks for a long time. Until recently, some researchers started creating benchmarks in the field of image fusion [124], [166]. Regarding VIF, Zhang et al. [167] proposed the first visible-infrared image fusion benchmark (VIFB), which consists of a test set of 21 visible-infrared image pairs, a code library of 20 VIF methods and 13 evaluation metrics. VIFB has been adopted in many VIF studies [143], [159], [168], [169], [170], [171], [172], [173], [174].

H. Application-Oriented VIF Methods

Most existing VIF methods do not consider downstream applications in the image fusion process, as shown in Fig. 10(a). Therefore, the features learned and fused during the image fusion process are general features, which may lead to visually-pleasing fused images, but may not be optimal for downstream applications.

In the past three years, a very important characteristic has been developing application-oriented VIF methods. This is very different from most existing VIF studies that do not consider downstream applications in the image fusion process. To the best of our knowledge, Shopovska et al. [44] is one of the earliest studies to consider downstream applications in VIF. Specifically, an auxiliary pedestrian detection error was employed in the loss function to help define relevant features of human appearance. The main focus of that work is to enhance the visibility of pedestrians in the fused image for human observers. Recently, Tang et al. [108] proposed SeAFusion that considers scene segmentation in the image fusion process. Liu et al. [115] formulated image fusion and object detection in a bilevel optimization formulation and proposed a joint training strategy to train the fusion model and the detection model together. Peng et al. [139] also used an image fusion loss term and an object detection term in the loss function. Compared to most VIF methods, these methods directly consider the performance of downstream applications in the fusion process by including application-based terms in the loss function, as shown in Fig. 10(b). Therefore, these methods can provide fused images that are more suitable for specific applications.

I. Different Terms in the Loss Function

A common characteristic of deep learning-based VIF methods is that different terms are usually included in the loss function. A primary reason of this is because there is no ground truth, thus the fusion quality is greatly dependent on the loss function. Consequently, researchers have to design their own loss functions to guide model training. When designing loss functions, a straightforward way is to design the loss function according to image fusion evaluation metrics. Indeed, almost all deep learning-based VIF methods contain loss terms designed according to image fusion evaluation metrics. Note that the loss function of most deep learning-based VIF methods only consider image fusion performance. Therefore, we call this kind of loss function as *VIF loss*, as shown in Fig. 10(a). However, as demonstrated by Zhang et al. [167], a VIF method may have very different performance in terms of different types of image fusion evaluation metrics, such as the structure-based metrics and information theory-based metrics. Consequently, a single metric-based VIF loss is not adequate to train a good VIF method. Therefore, researchers started to include different terms in the VIF loss. Usually, some of these terms correspond to different types of evaluation metrics. For example, Li et al. [117] use an intensity loss term and a structural similarity index measure (SSIM) loss term, Tang et al. [118] use a spatial frequency (SF) loss term and a SSIM loss term, Cheng et al. [175] use a SSIM loss term, a gradient-based loss term and a mean squared error (MSE) loss term.

Moreover, as mentioned in Section IV-H, researchers have begun to include application-based terms in the loss function in addition to VIF loss. We call these terms as *application loss*, as shown in Fig. 10(b). For example, Shopovska et al. [44] and Liu et al. [115] added object detection loss term to the loss function, and Tang et al. [108] added semantic segmentation loss term to the loss function.

In summary, most existing VIF methods only use VIF loss. However, a more promising way is to use both VIF loss and application loss. It is worth mentioning that the application loss is usually computed between the network output and ground truth of applications. In contrast, VIF loss is usually computed between the fused image and source images or pseudo ground truth.

J. Methods That Can Fuse Color Images Directly

Most VIF methods can only fuse grayscale images. To fuse color images, these methods first convert RGB images to the YCbCr space and then fuse the Y channel with the infrared image [91], [116], [118], [121]. An inverse color space transformation is then applied to obtain the color fused image. However, this process is complex. Moreover, most methods only fuse the Y channel using deep learning methods, while fuse Cr and

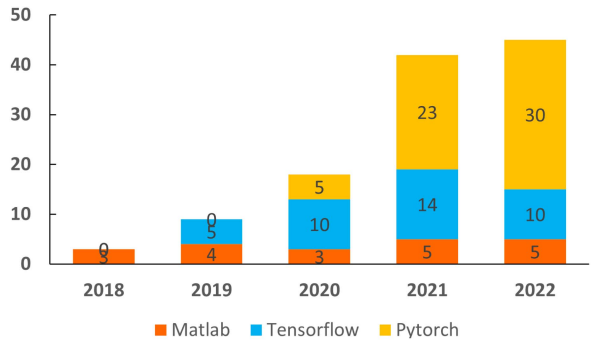


Fig. 11. The number of deep learning-based VIF methods (including general image fusion methods that can be applied to VIF) using different programming frameworks. We only countered papers that provide open-source code or explicitly mention the used programming framework.

Cb channels using conventional methods, e.g., manual methods. This may lead to information loss because Cb and Cr channels also contain important information. Recently, researchers proposed some VIF methods [25] that can fuse visible color and infrared images directly, which we believe is an important development characteristic and a promising future trend.

K. Programming Frameworks

We reviewed all the deep learning-based VIF methods shown in Fig. 2 to check the used programming framework. Fig. 11 shows the programming framework used in each year. As can be seen, the number of methods using Tensorflow increased quickly until 2021 and starts to decrease in 2022. In contrast, the number of methods using Pytorch increases very fast, indicating Pytorch has become the most popular programming framework used in deep learning-based VIF methods. The number of methods using Matlab increases very slightly from 2018 to 2022.

V. Datasets

A. Training Data for Deep Learning-Based Methods

Ground truth does not exist in the VIF task. Therefore, it is not straightforward to develop supervised VIF methods. This is reflected in Table I where most deep learning-based VIF methods are unsupervised methods. However, researchers have tried various methods to generate pseudo 'ground truth' or use some other forms of labels to perform supervised training. In this section, we discuss the training data of supervised and unsupervised methods.

1) Training Data for Supervised Methods: In this section, we summarize the measures taken by all types of supervised methods to generate training data.

The first way is to use fused images generated by other methods as ground truth. For example, Li et al. [24] use GFF [146] to generate labels. Lebedev et al. [22] generated ground truth images with Laplacian pyramid algorithm accompanied by MultiScale Retinex [144]. However, this method may set an upper limit for learning [101].

The second way is to use all-clear images and their blurred versions. Typical examples are [23], [29], [47], [56], [82], which uses RGB images and their blurred versions. Recently, Zhu et al. [112] generated blurred versions for both RGB and infrared images. However, the training data generated in this way is not very realistic and is different from real visible-infrared image pairs.

The third way is to use manually-labeled object masks for existing VIF dataset. For example, some researchers [31], [94], [95], [96] employ masks to help the fused images maintain semantic information. However, it is labor-intensive and not convenient to obtain these masks.

The fourth way is to use labels of downstream applications. In this way, the VIF task that does not have ground truth is converted to a task that has ground truth for a part of the loss function. For example, Shopovska et al. [44] utilized a pre-trained pedestrian detector to generate pedestrian labels and then used these labels to construct an auxiliary detection loss to perform a detection-guided training. Tang et al. [108] used scene segmentation as a downstream task and utilized a segmentation loss term in the loss function to guide training. The scene segmentation labels are manually labeled by the authors of the scene segmentation dataset. Another example is that Liu et al. [115] used general object detection as a downstream task and added an object detection loss term to the loss function. Object detection labels are provided in that work.

The final way is using the Y channel of RGB images in the YCbCr space as ground truth and generating synthetic infrared and visible images for training [102].

2) Training Data for Unsupervised Methods: There are several types of training data in unsupervised VIF methods. The first type is visible-infrared image pairs [37], [75], [143]. The second type is visible and infrared images, but they are not necessarily pairs. Some AE-based methods [28], [53], [72], [76], [78] using this kind of training data. The third type is all-clear visible images, which are mainly used in AE-based methods. For example, Li et al. [42] use the MS-COCO dataset to train the encoder and decoder. The fourth type is visible images plus visible-infrared image pairs. In this case, visible images and visible-infrared image pairs are used to train different modules of a model. For example, Jian et al. [90] use visible images to train an image decomposition module and visible-infrared image pairs to train a stacked sparse autoencoder for local saliency map extraction. Another example is the RFN-Nest method [73]. The fifth way to handle the lack of training data is transfer learning, i.e., using pre-trained models trained with large-scale RGB datasets, as introduced in Section II-D3.

We summarize the main methods of training data generation for both supervised and unsupervised methods in the supplementary material, (available online).

B. Test Sets

Some VIF datasets, including TNO [176], INO [177], MFNet [178], RoadScene [61], VIFB [167], LLVIP [179] and M3FD [115], have been utilized in the VIF task as test sets to evaluate image fusion performance. The information of these

TABLE II
SOME VIF METHODS PUBLISHED IN TOP JOURNALS AND CONFERENCES. AS CAN BE SEEN, IT IS DIFFICULT TO KNOW THE REAL PERFORMANCE COMPARISON OF VIF METHODS BECAUSE DIFFERENT TEST SETS AND EVALUATION METRICS ARE USED

Reference (year)	Venue	No. of test image pairs	Objective evaluation metrics
FusionGAN [32] (18)	INFUS	7 (TNO) + 31 (INO)	5 (EN, SD, SSIM, SF, VIF)
VIF-Net [59] (19)	TIP	9 (TNO and INO)	5 (MI, $Q^{AB/F}$, PC, Q^{NCIE}, UIQI)
U2Fusion [63] (19)	TPAMI	20 (TNO) + 45 (RoadScene)	4 (*SSIM, PSNR, CC, SCD*)
GANMcC [101] (20)	TIM	16 (TNO) + 30 (RoadScene)	6 (SSIM, CC, SCD, EN, SD, MI)
Liu et al. [92] (20)	TIP	20 (TNO)	5 (VIF, AG, SF, SCD, $Q^{AB/F}$)
SDNet [93] (20)	IJCV	10 (TNO)	4 (EN, FMI_{dct}, PSNR, MG)
Xu et al. [190] (22)	PR	20 (TNO) + 20 (RoadScene)	6 (Q_{abf}, SCD, VIF, SF, SSIM, PSNR)
MHTNet [114] (21)	TIM	20 (TNO) + 20 (KAIST) + 20 (BEMP)	6 (SCD, VIFF, EN, SD, MI, Q_{CV})
DDcGAN [49] (18)	IJCAI	20 (TNO)	4 (EN, SD, SF, PSNR)
DIDFuse [54] (19)	IJCAI	40 (TNO) + 52 (NIR) + 40 (FLIR)	6 (EN, MI, SD, SF, VIF, AG)
FusionDN [61] (19)	AAAI	44 (RoadScene)	4 (SD, EN, VIF, SCD)
PMGI [62] (19)	AAAI	17 (TNO)	6 (SSIM, $Q^{AB/F}$, EN, FMI, SCD, CC)
Liu et al. [91] (20)	MM	37 (TNO) + 26 (RoadScene)	4 (SD, VIF, CC, SCD)

datasets is summarized in the supplementary material, available online. It should be mentioned that some studies also use these datasets as training data. In general, the VIF field does not have a well-established test set, as shown in Table II. This is different from many other tasks in computer vision, e.g., object tracking and detection.

C. Other Datasets Containing Visible-Infrared Images

Apart from the VIF datasets mentioned above, some other datasets also contain visible and infrared images. For example, CVC-14 [180] and FLIR [181] provide visible-infrared image pairs for driving scenarios. However, the images in these datasets are not aligned. GTOT [182], RGBT234 [9], and LasHeR [183] are mainly used for RGBT tracking. They provide a large number of visible-infrared image pairs. However, the alignment of visible-infrared images in these datasets is not very accurate. In addition, multispectral KAIST [184] is a multispectral dataset that is mainly used for multispectral pedestrian detection. Moreover, the OSU dataset [185] is an earlier dataset used in VIF.

VI. PERFORMANCE EVALUATION METHODS

This section summarizes performance evaluation methods used in VIF. It is not an easy task to perform performance evaluation for VIF methods due to the lack of ground truth. Generally speaking, qualitative evaluation for visual performance and quantitative evaluation based on image fusion evaluation metrics are employed in existing studies.

A. Qualitative Evaluation

Qualitative evaluation means that the quality of fused images is checked manually and visually. Usually, the fused image should contain texture details of the visible image and salient features of the infrared image. Qualitative evaluation is crucial for the performance evaluation of VIF methods and has been chosen in almost all VIF papers. However, qualitative evaluation is not automatic hence is time-consuming. Moreover, it is not feasible to manually check all fused images, especially when a large number of fused images are generated. A common way in the existing literature is to select several examples for comparison. This may lead to sampling bias [186], which is a common issue in image enhancement tasks. Moreover, different observers may have different standards when checking fused images, resulting in subjective bias [186]. To the best of our knowledge, there is not a good solution for these issues currently.

B. Quantitative Evaluation

Quantitative evaluation means that the quality of fused images is checked using image fusion evaluation metrics. Many evaluation metrics, such as cross entropy (CE) [187], spatial frequency (SF) [188], and normalized mutual information (NMI) [189], have been proposed. However, there is not a well-recognized metric that has been used in most VIF studies. Moreover, each metric usually partially evaluates the quality of fused images from one aspect or very limited aspects. This results in a significant issue that different metrics may be used in different VIF studies, as shown in Table II. Moreover, in existing VIF literature, different test sets are also used. Therefore, it is pretty difficult to compare the performance of VIF methods fairly.

Recently, Zhang et al. [167] proposed the first VIF benchmark (VIFB), which can be utilized to perform comprehensive performance comparisons for VIF methods. VIFB contains 21 visible-infrared image pairs, 20 VIF methods, and 13 evaluation metrics. All fusion results are publicly available. VIFB is an important step towards developing better ways to evaluate image fusion methods.

Performance evaluation of VIF methods is an active topic. Some researchers are working on proposing better evaluation metrics or evaluation methods. More details about evaluation metrics can be found in [19], [123], [191].

VII. FUTURE PROSPECTS

In this section we discuss future prospects of VIF.

A. Better Evaluation Metrics

VIF methods are evaluated by qualitative comparison using visual performance and quantitative comparison using image fusion evaluation metrics in existing papers. However, as shown in Table II, it is prevalent that different metrics and test images are selected in different papers, making the fair performance comparison difficult. Moreover, one metric usually only evaluates the fusion results from certain aspects [167]. Furthermore, qualitative results are usually not consistent with quantitative results in the image fusion field [167], [192]. It is, therefore, desirable to have better metrics. Ideal metrics should be consistent with visual performance and reflect fusion performance comprehensively.

B. Better Benchmarks

Although Zhang et al. [167] developed a VIF benchmark, the research on VIF benchmark is still at an early stage. We selected five recent deep learning-based methods (IFCNN [23], SeAFusion [108], SwinFusion [121], U2Fusion [63], and YDTR [118]) that can represent the latest development in the VIF field and tested their performance on VIFB [167]. In these

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON ON VIFB. THE BEST THREE VALUES OF EACH METRIC ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY. THE THREE NUMBERS AFTER METHOD NAMES DENOTE THE NUMBER OF BEST VALUE, SECOND BEST VALUE, AND THIRD BEST VALUE, RESPECTIVELY. BEST VIEWED IN COLOR

Method	Information theory-based				Information feature-based					Structural similarity-based		Human perception-inspired	
	CE (↓)	EN (↑)	MI (↑)	PSNR (↑)	AG (↑)	EI (↑)	$Q^{AB/F}$ (↑)	SD (↑)	SF (↑)	RMSE (↓)	SSIM (↑)	Q_{CB} (↑)	Q_{CV} (↓)
ADF (0,0,0)	1.464	6.788	1.921	58.405	4.582	46.529	0.519	35.185	14.132	0.104	1.400	0.474	777.817
CBF (0,0,2)	0.994	7.324	2.161	57.595	7.154	74.590	0.578	48.544	20.380	0.126	1.171	0.526	1575.148
FPDE (0,0,0)	1.366	6.766	1.924	58.402	4.538	46.022	0.484	34.931	13.468	0.104	1.387	0.460	780.114
GFCE (0,3,0)	1.931	7.266	1.844	55.939	7.498	77.465	0.471	51.563	22.463	0.173	1.134	0.535	898.946
GFF (0,0,0)	1.189	7.210	2.638	58.100	5.326	55.198	0.624	50.059	17.272	0.112	1.398	0.619	881.625
GTF (0,0,0)	1.285	6.508	1.991	57.861	4.303	43.664	0.439	35.130	14.743	0.118	1.371	0.414	2138.369
HMSD_GF (0,1,0)	1.164	7.274	2.472	57.940	6.246	65.034	0.623	57.617	19.904	0.116	1.394	0.604	532.958
Hybrid_MSD (0,1,0)	1.257	7.304	2.619	58.173	6.126	63.491	0.636	54.922	19.659	0.110	1.405	0.623	510.866
IFEVIP (0,0,0)	1.339	6.936	2.248	57.174	4.984	51.782	0.486	48.491	15.846	0.138	1.391	0.462	573.767
LatLRR (3,0,0)	1.684	6.909	1.653	56.180	8.962	92.813	0.438	57.133	29.537	0.169	1.184	0.497	697.286
LP_SR (2,2,2)	0.957	7.339	2.809	57.951	5.851	60.781	0.661	57.314	18.807	0.117	1.390	0.645	522.687
MGFF (0,0,0)	1.295	7.114	1.768	58.212	5.839	60.607	0.573	44.290	17.916	0.109	1.406	0.542	676.887
MSVD (0,0,2)	1.462	6.705	1.955	58.415	3.545	36.202	0.331	34.372	12.525	0.104	1.425	0.426	808.993
NSCT_SR (3,0,1)	0.900	7.396	2.988	57.435	6.492	67.956	0.646	52.475	19.389	0.131	1.277	0.617	1447.340
RP_SR (0,1,2)	0.994	7.353	2.336	57.777	6.364	65.220	0.566	55.808	21.171	0.122	1.332	0.606	888.848
TIF (0,0,0)	1.371	7.075	1.767	58.225	5.558	57.839	0.584	42.643	17.739	0.109	1.399	0.545	613.004
VSMWLS (0,0,0)	1.409	7.028	2.035	58.194	5.612	57.252	0.554	46.253	17.662	0.109	1.417	0.496	754.704
CNN (1,1,2)	1.030	7.320	2.653	57.932	5.808	60.241	0.658	60.075	18.813	0.118	1.391	0.621	512.569
DLF (3,0,0)	1.413	6.724	2.030	58.444	3.825	38.569	0.434	34.717	12.491	0.103	1.461	0.445	759.814
IFCNN (0,0,1)	1.419	7.122	2.068	58.246	6.228	64.645	0.589	48.521	19.359	0.108	1.403	0.531	495.289
ResNet (0,3,0)	1.364	6.734	1.988	58.441	3.674	37.255	0.407	34.940	11.736	0.104	1.460	0.445	724.831
SeAFusion (0,1,0)	1.543	6.967	2.120	57.301	5.655	58.877	0.561	49.628	17.733	0.134	1.393	0.460	416.935
SwinFusion (1,0,0)	1.338	6.938	2.282	57.321	5.605	57.992	0.575	52.855	18.045	0.135	1.406	0.489	399.224
U2Fusion (0,0,0)	1.316	7.200	1.946	57.966	6.241	65.831	0.532	50.058	18.288	0.114	1.331	0.540	719.791
YDTR (0,0,1)	1.568	6.828	2.124	58.015	4.333	44.591	0.452	44.980	15.082	0.116	1.435	0.436	679.953

methods, IFCNN and U2Fusion are CNN-based general image fusion methods, SeAFusion is an application-driven VIF method, SwinFusion is a transformer-based general image fusion method, YDTR is a transformer-based VIF method. Quantitative results can be found in Table III. Due to the page limitation, we put qualitative results in the supplementary material, available online. We have several observations from the results on VIFB. First, the lastest deep learning-based methods do not show advantages over older deep learning-based methods on VIFB. For example, the three methods published in 2022, i.e., SeAFusion, SwinFusion, and YDTR, show worse quantitative performance than CNN and DLF on VIFB. Second, some conventional VIF methods, i.e., LatLRR, LP_SR and NSCT_SR, show very competitive quantitative performance with deep learning-based ones, indicating that deep learning-based methods do not exhibit dominant performance on VIFB. Third, it is hard to conclude which type of method is better from the qualitative comparison. Specifically, the latest deep learning-based methods show good fusion performance in some cases but give bad performance in other cases. Some conventional VIF methods also show very competitive qualitative performance. Finally, quantitative performance is not very consistent with qualitative performance, which is a common issue in the image fusion field. Note that these observations are based on VIFB [167]. Different observations may be obtained if a different set of test images and evaluation metrics are used.

VIFB is an initial effort in the development of VIF benchmarks. It has some limitations, e.g., the small number and low resolution of test images. More efforts are required to develop better benchmarks to compare VIF methods in a better way. For example, future benchmarks may contain more test images, more diverse scenarios, and a better combination of high-quality evaluation metrics.

C. Transformer-Based Methods

Transformers have achieved excellent performance in many computer vision tasks. However, as introduced in Section II-G, the application of transformers in VIF is at a very early stage. We expect that many transformer-based VIF methods will emerge in the coming years. In particular, it is interesting to develop pure transformer-based VIF methods. Moreover, it is essential to demonstrate what is global information in the context of VIF, which has been rarely explained in existing transformer-based VIF methods.

D. Application-Oriented Image Fusion Methods

One of the motivations for using VIF is to improve the performance of downstream applications. However, it can be seen from our review that most existing VIF methods do not take downstream applications into account. This can also be observed from the performance evaluation methods, i.e., qualitative comparison using visual performance and quantitative comparison using image fusion metrics. However, VIF methods designed in this way learn general features and fusion rules that may not be optimized for downstream applications. Therefore, it is better to consider the downstream applications in the design of VIF methods. A possible framework is shown in Fig. 10(b), where both VIF loss and application loss are used to guide training. We expect that application-oriented image fusion methods will become mainstream methods in this field.

E. More Applications

VIF has the potential to improve the performance of many applications, especially those that need to work under various illumination conditions. However, VIF has been mainly applied to object tracking [7], [193], object detection [14], [115], salient

object detection [31], [194], and scene segmentation [108], [195]. Many other applications, such as people rescuing [196] and robotics [197], are of great values but have been rarely investigated. We think more applications of VIF should be explored in the future.

F. Handling of Misalignment

The misalignment of visible and infrared images may degrade performance of applications, e.g., pedestrian detection [198]. It is thus very important to handle misalignment for fusion. This will also be helpful to promote the applications of VIF methods. However, although many studies have been conducted to handle misalignment of visible and infrared images, alignment is still an open question and it is very challenging to align visible and infrared images perfectly. Indeed, almost all existing RGB-infrared datasets, such as LasHeR [183], RGBT234 [9] and M^3FD [115], have some issues of misalignment. Deep learning might provide potential solutions to this problem, as explored by two recent studies [111], [164] which learn image fusion and registration together. It is also worth investigating image fusion, registration, and downstream applications together, as done by Tang et al. [199]. We expect that more deep learning-based registration methods will be proposed in the future to solve the misalignment issue.

G. Combining VIF With Other Tasks

In most VIF studies, only visible and infrared image fusion has been considered. Recently, some researchers performed VIF and other tasks together, which might be more effective and efficient. For example, Li et al. [27] and Gu et al. [159] combined VIF with image super-resolution. However, the research on the combination of VIF with other tasks is still very limited. We expect that more studies along this direction will be proposed in the coming years to further explore the mutual benefits of VIF and other tasks.

H. Improving Fusion Efficiency

With the development of deep learning-based VIF methods, researchers have designed larger and deeper models to perform VIF. However, larger models make the VIF method not efficient enough, which hinders the values of VIF methods in real applications, such as object tracking and detection. Some researchers [200] have noticed this and tried to design efficient deep VIF methods. However, these studies are still very limited. Designing efficient VIF methods will be an important trend of VIF in the future.

VIII. Conclusion

In this paper, a detailed review of deep learning-based visible and infrared image fusion (VIF) methods is presented. From the review, one can see that an increasing number of deep learning-based VIF methods are developed every year since 2018, and various deep learning techniques have been applied to perform VIF. We carefully grouped existing methods and introduced representative methods. We also discussed recent development characteristics of this field. Moreover, we summarized VIF datasets, including test data and training data, and performance evaluation methods. Based on these reviews and analysis, we discussed future prospects of VIF by analyzing several important issues that we think should attract more attention. We expect that this study can serve as a suitable reference for researchers in this field.

References

[1] A. Toet, L. J. Van Ruyven, and J. M. Valeton, "Merging thermal and visual images by a contrast pyramid," *Opt. Eng.*, vol. 28, no. 7, pp. 789–792, 1989.

[2] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, 2015.

[3] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, 2016.

[4] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infrared Phys. Technol.*, vol. 83, pp. 227–237, 2017.

[5] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process*, vol. 29, pp. 4733–4746, 2020.

[6] J. Ma et al., "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, 2020.

[7] X. Zhang, G. Xiao, P. Ye, D. Qiao, J. Zhao, and S. Peng, "Object fusion tracking based on visible and infrared images using fully convolutional siamese networks," in *Proc. IEEE 22nd Int. Conf. Inf. Fusion*, 2019, pp. 1–8.

[8] X. Zhang et al., "DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion," *Signal Process. Image Commun.*, vol. 84, 2020, Art. no. 115756.

[9] C. Li et al., "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, 2019, Art. no. 106977.

[10] V. Chandrakanth, M. V., and S. Channappayya, "Siamese cross domain tracker design for seamless tracking of targets in RGB and thermal videos," *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 161–172, Feb. 2023.

[11] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal UAV tracking: A large-scale benchmark and new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8886–8895.

[12] Y. Yan et al., "Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos," *Cogn. Comput.*, vol. 10, no. 1, pp. 94–104, 2018.

[13] R. Lahmyed, M. El Ansari, and A. Ellahyani, "A new thermal infrared and visible spectrum images-based pedestrian detection system," *Multimedia Tools Appl.*, vol. 78, no. 12, pp. 15 861–15 885, 2019.

[14] H. Sun et al., "Fusion of infrared and visible images for remote detection of low-altitude slow-speed small targets," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2971–2983, 2021.

[15] H. Zhou et al., "Visible-thermal image object detection via the combination of illumination conditions and temperature information," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3656.

[16] S. G. Kong et al., "Recent advances in visual and infrared face recognition - A review," *Comput. Vis. Image Understanding*, vol. 97, no. 1, pp. 103–135, 2005.

[17] S. Ariffin, N. Jamil, and P. Rahman, "Can thermal and visible image fusion improves ear recognition?," in *Proc. 8th Int. Conf. Inf. Technol.*, 2017, pp. 780–784.

[18] D. P. Bavirisetti et al., "Multi-scale guided image and video fusion: A fast and efficient approach," *Circuits, Syst. Signal Process.*, vol. 38, no. 12, pp. 5576–5605, 2019.

[19] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.

[20] Y. Liu et al., "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, 2018.

[21] W. Wu et al., "Visible and infrared image fusion using NSST and deep Boltzmann machine," *Optik*, vol. 157, pp. 334–342, 2018.
[22] M. Lebedev, D. Komarov, O. Vygolov, and Y. V. Vizilter, "Multisensor image fusion based on generative adversarial networks," in *Image and Signal Processing for Remote Sensing XXV*, vol. 11155. Bellingham, WA, USA: SPIE, 2019, pp. 565–574.
[23] Y. Zhang et al., "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.
[24] Q. Li et al., "Coupled GAN with relativistic discriminators for infrared and visible images fusion," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7458–7467, Mar. 2021.
[25] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Trans. Image Process*, vol. 29, pp. 3845–3858, 2020.
[26] J. Wang et al., "Infrared and visible image fusion based on Laplacian pyramid and generative adversarial network," *KSII Trans. Internet Inf. Syst.*, vol. 15, no. 5, pp. 1761–1777, 2021.
[27] H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," *IEEE Trans. Image Process*, vol. 30, pp. 4070–4083, 2021.
[28] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDRFuse: A symmetric encoder-decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.
[29] Y. Liu et al., "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 03, 2018, Art. no. 1850018.
[30] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2705–2710.
[31] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[32] J. Ma et al., "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, 2019.
[33] Y. Yang, J. Liu, S. Huang, W. Wan, W. Wen, and J. Guan, "Infrared and visible image fusion via texture conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4771–4783, Dec. 2021.
[34] V. Vs, J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 3566–3570.
[35] H. Zhao and R. Nie, "DNDT: Infrared and visible image fusion via DenseNet and dual-transformer," in *Proc. IEEE Int. Conf. Inf. Technol. Biomed. Eng.*, 2021, pp. 71–75.
[36] Y. Fu, T. Xu, X. Wu, and J. Kittler, "PPT Fusion: Pyramid patch transformer for a case study in image fusion," 2021, *arXiv:2107.13967*.
[37] D. Rao, X.-J. Wu, and T. Xu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," 2022, *arXiv:2201.10147*.
[38] X. Zhang et al., "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, vol. 63, pp. 166–187, 2020.
[39] C. Sun, C. Zhang, and N. Xiong, "Infrared and visible image fusion techniques based on deep learning: A review," *Electronics*, vol. 9, no. 12, 2020, Art. no. 2162.
[40] H. Zhang et al., "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.
[41] X. Ren et al., "Infrared-visible image fusion based on convolutional neural networks," in *Proc. Int. Conf. Intell. Sci. Big Data Eng.*, Springer, 2018, pp. 301–307.
[42] H. Li and X. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[43] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infrared Phys. Technol.*, vol. 102, 2019, Art. no. 103039.
[44] I. Shopovska, L. Jovanov, and W. Philips, "Deep visible and thermal image fusion for enhanced pedestrian visibility," *Sensors*, vol. 19, no. 17, 2019, Art. no. 3727.
[45] Y. Cui, H. Du, and W. Mei, "Infrared and visible image fusion using detail enhanced channel attention network," *IEEE Access*, vol. 7, pp. 182 185–182 197, 2019.
[46] Y. Liu et al., "Infrared and visible image fusion through details preservation," *Sensors*, vol. 19, no. 20, 2019, Art. no. 4556.
[47] M. Wang, X. Liu, and H. Jin, "A generative image fusion approach based on supervised deep convolution network driven by weighted gradient flow," *Image Vis. Comput.*, vol. 86, pp. 1–16, 2019.
[48] F. Lahoud and S. Süsstrunk, "Fast and efficient zero-learning image fusion," 2019, *arXiv: 1905.03590*.
[49] H. Xu et al., "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3954–3960.
[50] J. Li, H. Huo, K. Liu, C. Li, S. Li, and X. Yang, "Infrared and visible image fusion via multi-discriminators wasserstein generative adversarial network," in *Proc. IEEE 18th Int. Conf. Mach. Learn. Appl.*, 2019, pp. 2014–2019.
[51] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
[52] J. Zhu et al., "Multiscale channel attention network for infrared and visible image fusion," *Concurrency Comput. Pract. Experience*, vol. 33, 2021, Art. no. e6155.
[53] H. Patel et al., "An approach for fusion of thermal and visible images," in *Proc. Int. Conf. Emerg. Technol. Trends Electron. Commun. Netw.*, Springer, 2020, pp. 225–234.
[54] Z. Zhao et al., "DIDFuse: Deep image decomposition for infrared and visible image fusion," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 970–976.
[55] A. Raza, H. Hong, and T. Fang, "PFAF-Net: Pyramid feature network for multimodal fusion," *IEEE Sens. Lett.*, vol. 4, no. 12, pp. 1–4, Dec. 2020.
[56] Y. Feng et al., "Fully convolutional network-based infrared and visible image fusion," *Multimedia Tools Appl.*, vol. 79, no. 21, pp. 15 001–15 014, 2020.
[57] Y. Li et al., "Unsupervised densely attention network for infrared and visible image fusion," *Multimedia Tools Appl.*, vol. 79, no. 45, pp. 34 685–34 696, 2020.
[58] W.-B. An and H.-M. Wang, "Infrared and visible image fusion with supervised convolutional neural network," *Optik*, vol. 219, 2020, Art. no. 165120.
[59] R. Hou et al., "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 640–651, 2020.
[60] H. T. Mustafa et al., "Infrared and visible image fusion based on dilated residual attention network," *Optik*, vol. 224, 2020, Art. no. 165409.
[61] H. Xu et al., "FusionDN: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12484–12491.
[62] H. Zhang et al., "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12797–12804.
[63] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
[64] F. Zhao and W. Zhao, "Learning specific and general realm feature representations for image fusion," *IEEE Trans. Multimedia*, vol. 23, pp. 2745–2756, 2021.
[65] D. Xu et al., "Infrared and visible image fusion with a generative adversarial network and a residual network," *Appl. Sci.*, vol. 10, no. 2, 2020, Art. no. 554.
[66] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.
[67] J. Li et al., "Infrared and visible image fusion using dual discriminators generative adversarial networks with wasserstein distance," *Inf. Sci.*, vol. 529, pp. 28–41, 2020.
[68] J. Xu et al., "LBP-BEGAN: A generative adversarial network architecture for infrared and visible image fusion," *Infrared Phys. Technol.*, vol. 104, 2020, Art. no. 103144.
[69] C. Yuan et al., "FLGC-Fusion GAN: An enhanced fusion GAN model by importing fully learnable group convolution," *Math. Problems Eng.*, vol. 2020, pp. 1–13, 2020.
[70] Y. Zhao, G. Fu, H. Wang, and S. Zhang, "The fusion of unmatched infrared and visible images based on generative adversarial networks," *Math. Problems Eng.*, vol. 2020, pp. 1–12, 2020.

[71] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process*, vol. 29, pp. 4980–4995, 2020.

[72] Y. Fu and X.-J. Wu, "A dual-branch network for infrared and visible image fusion," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 10 675–10 680.

[73] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, 2021.

[74] Y. Fu, X.-J. Wu, and J. Kittler, "Effective method for fusing infrared and visible images," *J. Electron. Imag.*, vol. 30, no. 3, 2021, Art. no. 033013.

[75] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 824–836, 2021.

[76] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1186–1196, Mar. 2022.

[77] F. Zhao, W. Zhao, L. Yao, and Y. Liu, "Self-supervised feature adaption for infrared and visible image fusion," *Inf. Fusion*, vol. 76, pp. 189–203, 2021.

[78] H. Patel and K. P. Upla, "DepthFuseNet: An approach for fusion of thermal and visible images using a convolutional neural network," *Opt. Eng.*, vol. 60, no. 1, 2021, Art. no. 013104.

[79] Y. Pan et al., "DenseNetFuse: A study of deep unsupervised DenseNet to infrared and visual image fusion," *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 10339–10351, 2021.

[80] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3360–3374, Jun. 2022.

[81] J. Liu, Y. Wu, Z. Huang, R. Liu, and X. Fan, "SMoA: Searching a modality-oriented architecture for infrared and visible image fusion," *IEEE Signal Process. Lett.*, vol. 28, pp. 1818–1822, 2021.

[82] X. Luo, Y. Gao, A. Wang, Z. Zhang, and X. -J. Wu, "IFSepR: A general framework for image fusion based on separate representation learning," *IEEE Trans. Multimedia*, vol. 25, pp. 608–623, 2023.

[83] Y. Long et al., "RXDNFuse: A aggregated residual dense network for infrared and visible image fusion," *Inf. Fusion*, vol. 69, pp. 128–141, 2021.

[84] Y. Yang et al., "VMDM-fusion: A saliency feature representation method for infrared and visible image fusion," *Signal Image Video Process.*, vol. 15, no. 6, pp. 1221–1229, 2021.

[85] K. Ren et al., "An infrared and visible image fusion method based on improved DenseNet and mRMR-ZCA," *Infrared Phys. Technol.*, vol. 115, 2021, Art. no. 103707.

[86] A. Raza et al., "IR-MSDNet: Infrared and visible image fusion based on infrared features multiscale dense network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3426–3437, 2021.

[87] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[88] Z. Xu, G. Liu, L. L. Tang, and Y. H. Li, "Blur regional features based infrared and visible image fusion using an improved C3Net model," *J. Phys. Conf. Ser.*, vol. 1820, no. 1. IOP Publishing, 2021, Art. no. 012169.

[89] L. Liu, M. Chen, M. Xu, and X. Li, "Two-stream network for infrared and visible images fusion," *Neurocomputing*, vol. 460, pp. 50–58, 2021.

[90] L. Jian, R. Rayhana, L. Ma, S. Wu, Z. Liu, and H. Jiang, "Infrared and visible image fusion based on deep decomposition network and saliency analysis," *IEEE Trans. Multimedia*, vol. 24, pp. 3314–3326, 2021.

[91] R. Liu et al., "Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1600–1608.

[92] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE Trans. Image Process*, vol. 30, pp. 1261–1274, 2021.

[93] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, pp. 2761–2785, 2021.

[94] Y. Gu, X. Wang, C. Zhang, and B. Li, "Advanced driving assistance based on the fusion of infrared and visible images," *Entropy*, vol. 23, no. 2, 2021, Art. no. 239.

[95] J. Hou et al., "A generative adversarial network for infrared and visible image fusion based on semantic segmentation," *Entropy*, vol. 23, no. 3, 2021, Art. no. 376.

[96] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, "Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network," *IEEE Trans. Multimedia*, vol. 25, pp. 635–648, 2023.

[97] S. Bhagat, S. D. Joshi, B. Lall, and S. Gupta, "Multimodal sensor fusion using symmetric skip autoencoder via an adversarial regulariser," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1146–1157, 2021.

[98] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[99] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.

[100] Y. Fu, X.-J. Wu, and T. Durrani, "Image fusion based on generative adversarial network consistent with perception," *Inf. Fusion*, vol. 72, pp. 110–125, 2021.

[101] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[102] X. Luo, A. Wang, Z. Zhang, X. Xiang, and X. -J. Wu, "LatRAIVF: An infrared and visible image fusion method based on latent regression and adversarial training," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–16, 2021.

[103] H. Zhang, J. Yuan, X. Tian, and J. Ma, "GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1134–1147, 2021.

[104] X. Li, H. Chen, Y. Li, and Y. Peng, "MAFusion: Multiscale attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

[105] W. Su, Y. Huang, Q. Li, F. Zuo, and L. Liu, "Infrared and visible image fusion based on adversarial feature extraction and stable image reconstruction," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[106] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, "Res2Fusion: Infrared and visible image fusion based on dense Res2net and double nonlocal attention models," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[107] H. Xu et al., "CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Comput. Vis. Image Understanding*, vol. 218, 2022, Art. no. 103407.

[108] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, 2022.

[109] Z. Zhu, X. Yang, R. Lu, T. Shen, X. Xie, and T. Zhang, "CLF-Net: Contrastive learning for infrared and visible image fusion network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[110] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83, pp. 79–92, 2022.

[111] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3508–3515.

[112] D. Zhu, W. Zhan, Y. Jiang, X. Xu, and R. Guo, "IPLF: A novel image pair learning fusion network for infrared and visible image," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8808–8817, May 2022.

[113] C. Cheng et al., "StyleFuse: An unsupervised network based on style loss function for infrared and visible image fusion," *Signal Process. Image Commun.*, vol. 106, 2022, Art. no. 116722.

[114] Q. Li et al., "A multilevel hybrid transmission network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[115] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5792–5801.

[116] Y. Gao, S. Ma, and J. Liu, "DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 549–561, Feb. 2023.

[117] J. Li, J. Zhu, C. Li, X. Chen, and B. Yang, "CGTF: Convolution-guided transformer for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.
[118] W. Tang, F. He, and Y. Liu, "YDTR: Infrared and visible image fusion via y-shape dynamic transformer," *IEEE Trans. Multimedia*, early access, Jul. 20, 2022, doi: 10.1109/TMM.2022.3192661.
[119] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual swin transformer fusion network for infrared and visible images," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.
[120] X. Liu et al., "MFST: Multi-modal feature self-adaptive transformer for infrared and visible image fusion," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3233.
[121] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.
[122] G. Xiao, D. P. Bavirisetti, G. Liu, and X. Zhang, *Image Fusion*. Shanghai, China: Spinger Press & Shanghai Jiao Tong University Press, 2020.
[123] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4819–4838, Sep. 2022.
[124] X. Zhang, "Benchmarking and comparing multi-exposure image fusion algorithms," *Inf. Fusion*, vol. 74, pp. 111–131, 2021.
[125] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 19, pp. 4–19, 2014.
[126] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, 2016.
[127] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[128] D. Xu et al., "Multi-scale unsupervised network for infrared and visible image fusion based on joint attention mechanism," *Infrared Phys. Technol.*, vol. 125, 2022, Art. no. 104242.
[129] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[130] Z. Yang and S. Zeng, "TPFusion: Texture preserving fusion of infrared and visible images via dense networks," *Entropy*, vol. 24, no. 2, 2022, Art. no. 294.
[131] Z. Ding et al., "A robust infrared and visible image fusion framework via multi-receptive-field attention and color visual perception," *Appl. Intell.*, vol. 53, pp. 8114–8132, 2023.
[132] Y. Liu, C. Miao, J. Ji, and X. Li, "MMF: A Multi-scale MobileNet based fusion method for infrared and visible image," *Infrared Phys. Technol.*, vol. 119, 2021, Art. no. 103894.
[133] L. Yan, J. Cao, S. Rizvi, K. Zhang, Q. Hao, and X. Cheng, "Improving the performance of image fusion based on visual saliency weight map combined with CNN," *IEEE Access*, vol. 8, pp. 59 976–59 986, 2020.
[134] J. Liu, Y. Wu, G. Wu, R. Liu, and X. Fan, "Learn to search a lightweight architecture for target-aware infrared and visible image fusion," *IEEE Signal Process. Lett.*, vol. 29, pp. 1614–1618, 2022.
[135] Y. Zou et al., "Infrared visible color night vision image fusion based on deep learning," in *AI and Optical Data Sciences II*, vol. 11703. Bellingham, WA, USA: SPIE, 2021, Art. no. 117031S.
[136] Z. Shen et al., "Cross attention-guided dense network for images fusion," 2021, *arXiv:2109.11393*.
[137] D. Zhang et al., "An infrared and visible image fusion method based on deep learning," in *Proc. 4th Opt. Young Scientist Summit*, International Society for Optics and Photonics, 2021, Art. no. *1178109*.
[138] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 740–755.
[139] Y. Peng et al., "MFDetection: A highly generalized object detection network unified with multilevel heterogeneous image fusion," *Optik*, vol. 266, 2022, Art. no. 169599.
[140] Z. Li et al., "Infrared and visible fusion imaging via double-layer fusion denoising neural network," *Digit. Signal Process.*, vol. 123, 2022, Art. no. 103433.
[141] J. Wang, Y. Li, and Z. Miao, "A new infrared and visible image fusion method based on generative adversarial networks and attention mechanism," in *Proc. Int. Conf. Image Graph. Process.*, 2021, pp. 109–119.
[142] B. Liao, Y. Du, and X. Yin, "Fusion of infrared-visible images in UE-IoT for fault point detection based on GAN," *IEEE Access*, vol. 8, pp. 79 754–79 763, 2020.
[143] A. Song, H. Duan, H. Pei, and L. Ding, "Triple-discriminator generative adversarial network for infrared and visible image fusion," *Neurocomputing*, vol. 483, pp. 183–194, 2022.
[144] A. B. Petro, C. Sbert, and J.-M. Morel, "Multiscale retinex," *Image Process. On Line*, vol. 4, pp. 71–88, 2014.
[145] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 469–477.
[146] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.
[147] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[148] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[149] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
[150] X. Yang et al., "DGLT-Fusion: A decoupled global-local infrared and visible image fusion transformer," *Infrared Phys. Technol.*, vol. 128, 2023, Art. no. 104522.
[151] W. Tang, F. He, and Y. Liu, "TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation," *Pattern Recognit.*, vol. 137, 2023, Art. no. 109295.
[152] L. Qu et al., "Transfuse: A unified transformer-based image fusion framework using self-supervised learning," 2022, *arXiv:2201.07451*.
[153] Q. Zhou et al., "Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21 741–21 761, 2022.
[154] X. Jin et al., "An unsupervised multi-focus image fusion method based on Transformer and U-Net," *IET Image Process.*, vol. 17, pp. 733–746, 2022.
[155] W. Tang, F. He, Y. Liu, and Y. Duan, "MATR: Multimodal medical image fusion via multiscale adaptive transformer," *IEEE Trans. Image Process*, vol. 31, pp. 5134–5149, 2022.
[156] L. Qu et al., "TransMEF: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2126–2134.
[157] Z. Le et al., "UIFGAN: An unsupervised continual-learning generative adversarial network for unified image fusion," *Inf. Fusion*, vol. 88, pp. 305–318, 2022.
[158] H. Zhou, J. Hou, Y. Zhang, J. Ma, and H. Ling, "Unified gradient-and intensity-discriminator generative adversarial network for image fusion," *Inf. Fusion*, vol. 88, pp. 184–201, 2022.
[159] Y. Gu et al., "A dataset-free self-supervised disentangled learning method for adaptive infrared and visible images super-resolution fusion," 2021, *arXiv:2112.02869*.
[160] W. Xiao, Y. Zhang, H. Wang, F. Li, and H. Jin, "Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.
[161] J. H. Lee, Y. S. Kim, D. Lee, D. -G. Kang, and J. B. Ra, "Robust CCD and IR image registration using gradient-based statistical information," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 347–350, Apr. 2010.
[162] J. Han, E. J. Pauwels, and P. De Zeeuw, "Visible and infrared image registration in man-made environments employing hybrid visual features," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 42–51, 2013.
[163] C. Min, Y. Gu, Y. Li, and F. Yang, "Non-rigid infrared and visible image registration by enhanced affine transformation," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107377.
[164] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 679–19 688.
[165] C. Liu, B. Yang, X. Zhang, and L. Pang, "IBPNet: A multi-resolution and multi-modal image fusion network via iterative back-projection," *Appl. Intell.*, vol. 52, pp. 16185–16201, 2022.
[166] X. Zhang, "Multi-focus image fusion: A benchmark," 2020, *arXiv: 2002.03322*.
[167] X. Zhang, P. Ye, and G. Xiao, "VIFB: A visible and infrared image fusion benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 468–478.
[168] Y. Yuan et al., "Defogging technology based on dual-channel sensor information fusion of near-infrared and visible light," *J. Sensors*, vol. 2020, pp. 1–17, 2020.
[169] A. Fang et al., "Non-linear and selective fusion of cross-modal images," *Pattern Recognit.*, vol. 119, 2021, Art. no. 108042.
[170] C. Zhang, H. Hu, Y. Tai, L. Yun, and J. Zhang, "Trustworthy image fusion with deep learning for wireless applications," *Wirel. Commun. Mobile Comput.*, vol. 2021, pp. 1–9, 2021.

[171] F. C. Ataman and G. B. Akar, "Visible and infrared image fusion using encoder-decoder network," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1779–1783.
[172] A. Fang et al., "A light-weight, efficient, and general cross-modal image fusion network," *Neurocomputing*, vol. 463, pp. 198–211, 2021.
[173] Z. Wang and B. Sun, "Explicit and implicit models in infrared and visible image fusion," 2022, *arXiv:2206.09581*.
[174] X. Lin, G. Zhou, X. Tu, Y. Huang, and X. Ding, "Two-level consistency metric for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.
[175] C. Cheng, X.-J. Wu, T. Xu, and G. Chen, "UNIFusion: A lightweight unified image fusion network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.
[176] TNO Image Fusion Dataset. Accessed: Oct. 10, 2022. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
[177] Videos Analytics Dataset. Accessed: Oct. 10, 2022. [Online]. Available: https://www.ino.ca/en/technologies/video-analytics-dataset/videos/
[178] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2017, pp. 5108–5115.
[179] X. Jia et al., "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 3496–3504.
[180] A. González et al., "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, 2016, Art. no. 820.
[181] Teledyne FLIR ADAS Dataset. Accessed: Oct. 10, 2022. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form//#anchor29
[182] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.
[183] C. Li et al., "LasHeR: A large-scale high-diversity benchmark for RGBT tracking," *IEEE Trans. Image Process*, vol. 31, pp. 392–404, 2021.
[184] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.
[185] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Understanding*, vol. 106, no. 2–3, pp. 162–182, 2007.
[186] P. Cao, Z. Wang, and K. Ma, "Debiased subjective assessment of real-world image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 711–721.
[187] D. M. Bulanon, T. Burks, and V. Alchanatis, "Image fusion of visible and thermal images for fruit detection," *Biosyst. Eng.*, vol. 103, no. 1, pp. 12–22, 2009.
[188] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[189] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on'information measure for performance of image fusion'," *Electron. Lett.*, vol. 44, no. 18, pp. 1066–1067, 2008.
[190] M. Xu, L. Tang, H. Zhang, and J. Ma, "Infrared and visible image fusion via parallel scene and texture learning," *Pattern Recognit.*, vol. 132, 2022, Art. no. 108929.
[191] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.
[192] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process*, vol. 29, pp. 7203–7216, 2020.
[193] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust RGBT tracking," *IEEE Trans. Intell. Veh.*, vol. 6, no. 1, pp. 121–130, Mar. 2021.
[194] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," *IEEE Trans. Multimedia*, early access, May 03, 2022, doi: 10.1109/TMM.2022.3171688.
[195] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for RGB-thermal scene parsing," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3571–3579.
[196] D. C. Schedl, I. Kurmi, and O. Bimber, "An autonomous drone for search and rescue in forests using airborne optical sectioning," *Sci. Robot.*, vol. 6, no. 55, 2021, Art. no. eabg1188.
[197] L. Chen, L. Sun, T. Yang, L. Fan, K. Huang, and Z. Xuanyuan, "RGB-T SLAM: A flexible SLAM framework by combining appearance and thermal information," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 5682–5687.
[198] L. Zhang et al., "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5127–5137.
[199] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.
[200] S. Özer, M. Ege, and M. A. Özkanoglu, "SiameseFuse: A computationally efficient and a not-so-deep network to fuse visible and infrared images," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108712.

Xingchen Zhang (Member, IEEE) received the BSc and PhD degrees from the Huazhong University of Science and Technology and Queen Mary University of London in 2012 and 2018, respectively. He is currently a Marie Skłodowska-Curie Individual fellow with the Department of Electrical and Electronic Engineering, Imperial College London. Prior to this, he was a teaching fellow and research associate with the same department. His main research interests include image fusion, human motion/intention prediction, and object tracking. He is a recipient of the Best Paper Honourable Mention Award of the 9th Chinese Conference on Information fusion and a co-author of the book Image Fusion. He is a reviewer for UKRI Future Leaders Fellowship, EPSRC New Investigator Award, and EPSRC Open Fellowship.

Yiannis Demiris (Senior Member, IEEE) received the BSc (Hons.) degree in artificial intelligence and computer science and the PhD degree in intelligent robotics from the Department of Artificial Intelligence, University of Edinburgh, U.K., in 1994 and 1999, respectively. He is a professor with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., where he is the Royal Academy of Engineering Chair in Emerging Technologies, and the Head of the Personal Robotics Laboratory. His current research interests include human-robot interaction, machine learning, user modeling, and assistive robotics. He has published more than 220 journal and peer-reviewed conference papers in the above areas. He was a recipient of the Rector's Award for Teaching Excellence in 2012 and the FoE Award for Excellence in Engineering Education in 2012. He is a fellow of the Institution of Engineering and Technology (IET).