IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 November 2023, accepted 19 December 2023, date of publication 1 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348789

METHODS

Integrating Parallel Attention Mechanisms and Multi-Scale Features for Infrared and Visible Image Fusion

QIAN XU[1] AND YUAN ZHENG[2]
[1]School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310027, China
[2]School of Computer Science, Civil Aviation Flight University of China, Guanghan 618311, China

Corresponding author: Qian Xu (qian_xu3@zju.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62088101.

ABSTRACT Infrared and visible image fusion (IVIF) aims to synthesize images that capitalize on the strengths of both modalities. Addressing the common challenge in IVIF of preserving thermal radiation from infrared and textural details from visible images, we introduce AMFusionNet. AMFusionNet uniquely combines a multi-kernel convolution block (MKCBlock) with parallel spatial attention and channel attention modules (PSCNet), streamlining the feature extraction process. This integration enhances the model's ability to simultaneously capture essential details from both image types. Additionally, we incorporate a multi-scale structural similarity (MS-SSIM) loss function in our comprehensive loss function to further refine the detail preservation in the fused images. Our experimental evaluations on the TNO and FLIR datasets demonstrate that AMFusionNet achieves superior performance in both objective and subjective assessments compared to recent methods.

INDEX TERMS Infrared and visible image fusion, parallel attention mechanism, multi-kernel convolution, MS-SSIM.

I. INTRODUCTION

Image fusion is an important problem in the field of computer vision because it can provide more detailed and reliable information for the enhanced understanding and description of complex scenes [1], [2], [3], [4]. Among all image-fusion scenarios, IVIF is one of the most popular [5]. Infrared and visible images exhibited remarkable complementarity. For example, infrared images can capture the thermal radiation emitted from objects; however, they lack textural details. By contrast, visible images typically contain abundant structural information. However, visible images are highly susceptible to environmental conditions, such as heavy fog and low-light conditions. Owing to the complementary characteristics of infrared and visible images, it is worthwhile to fuse them into a composite image. This integration is beneficial for various downstream visual tasks, such as target detection [6], [7], tracking [8], pedestrian re-identification [9], and semantic segmentation [10].

In recent decades, many methods have been proposed to achieve high-quality fusion of infrared and visible images. These image fusion methods can be divided into two categories: pixel-based and deep learning (DL)-based.

Piexl-level image fusion techniques include multi-scale transform (MST)-based [11], [12], [13], saliency-based [14], [15], sparse representation-based [16], [17], optimization-based [18], [19], and hybrid methods [20]. Among these, MST-based methods have attracted considerable attention because of their high flexibility and excellent performance in image fusion. Specifically, MST-based methods use wavelet and multiscale pyramid transforms, which generally include three steps. First, the source images were decomposed into a series of sub-images with different spatial resolutions. Then, the subimages are fused at the corresponding levels using a predefined fusion stategy. Third, the corresponding inverse transform is applied to these fused sub-images to obtain

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

the final fused image. Although pixel-level-based methods have achieved commendable results, excessive manual intervention may lead to significant ghosting and noise in fused images. Furthermore, manually designed feature extractors require manually designed feature extractors not only require a substantial period of time but also are lack of generalization capability.

DL based methods have achieved remarkable results in various fields, such as object recognition [21], [22], object segmentation [23], [24], and object tracking [25], [26]. Owing to their excellent ability to represent features, DL-based methods are becoming mainstream in IVIF. According to different fusion architectures, DL-based methods can be roughly divided into three classes: convolutional neural network (CNN)-based methods [27], [28], [29] and generative adversarial network (GAN)-based methods [30], [31], [32]. Although existing DL-based methods have achieved satisfactory results in some typical scenes of image fusion, some issues still need to be resolved. Methods based on CNN or GAN consistently employ convolutional operations to capture features from images with restricted receptive fields, and these uniform convolutional processes also limit the capabilities for feature extraction and representation, thereby leading to the loss of crucial information during the information extraction process. Recently, some researchers have introduced attention mechanisms into these three network architectures; however, the forms of these attention mechanisms are implemented in a sequential manner [33] for IVIF, which results in weakened information interaction between the spatial attention module and channel attention module within the attention mechanism.

To address the problems mentioned above, particularly the challenge of balancing the emphasis on salient targets with the preservation of texture detail, we propose AMFusionNet, a fusion architecture that leverages the strengths of PSCNet and MKCBlock, and incorporates a novel loss function for enhanced performance. Specifically, the architecture integrates spatial attention (SA) and channel attention (CA) networks in a parallel manner and employs GELU as the activation function. This parallel structure enables more efficient information fusion across attention modules, thereby reducing information loss and enhancing the capability of the network to extract features within the IVIF framework. Furthermore, the incorporation of attention mechanisms guides the attention of the network to salient regions within the images, ensuring the maximal retention of thermal radiation information in the fused images. However, clear foreground targets are also important in evaluating fused images and detailed background information also helps people better understand the image. To retain more textural details, the MS-SSIM loss function to the total loss function. We use three fusion rules: average weight, L_1-norm, and mean-filter-operator. Consequently, the fused images are clear in the foreground objects and rich in background details. In summary, our main contributions are as follows:

- We propose a feature extraction layer that integrates three MKCBlocks in the decoder. This enhances the network's feature extraction capability, extends its receptive field, and allows multi-scale feature processing, capturing richer semantic features from images.
- MS-SSIM is introduced into the total loss function. MS-SSIM calculates a multi-scale feature map loss function to assess the similarity between fused images and their corresponding source images. This ensures that the resulting fused image contains richer textural details and distinct contours of the salient thermal targets.
- We propose a parallel attention mechanism architecture that minimizes information loss during feature extraction, while concurrently enhancing the integration of information between the channel and spatial attention branches.
- Extensive experiments demonstrated that our method outperforms other state-of-the-art fusion techniques in both subjective and objective assessments.

II. RELATED WORK

Based on the previous section, it is evident that the existing methodologies for IVIF can generally be classified into two main categories: pixel-level and DL-based methods.

A. PIXEL-LEVEL IMAGE FUSION ALGORITHMS

Pixel-level fusion techniques can be grouped based on their underlying principles, including MST-based [35], [36], saliency-based [37], and sparse representation-based methods [38].

MST-based methods are particularly important in image fusion. When using MST-based methods to fuse images, it is assumed that the images can be represented by multiple layers at varying resolutions. MST-based methods primarily include pyramid and wavelet transforms, such as Laplacian Pyramid (LP) [39] and Discrete Wavelet Decomposition (DWT) [40]. They then applied specific fusion rules to each layer before reconstructing the fused images by using the corresponding inverse transforms. By considering feature variations at different scales, this approach preserves the image details and texture information. Therefore, the quality and clarity of the fused image can be enhanced [18], [41].

Saliency-based approaches are used to capture and preserve critical features from both infrared and visible images within the fused image. These techniques not only improve the visibility of objects and scenes under low-contrast or low-light conditions but also strengthen the detection and recognition functions of computer vision systems [42], [43]. Saliency-based methods follow three main steps. First, they generated saliency maps for infrared and visible images. Subsequently, the pertinent features were extracted from each image. Finally, these methodologies employ a specific fusion rule to integrate features, producing a single image that succinctly embodies the most significant elements of

FIGURE 1. From left to right: infrared image, visible image, the fusion result of FusionGAN [30], the fusion result of U2Fusion [34], and the fusion result of AMFusionNet.

both sources. The extracted key attributes typically comprise the following: Texture, edges, and other distinctive visual features that are essential for differentiating between various regions.

In the IVIF domain, sparse representation is widely employed as a prominent method. It seeks to derive an overcomplete dictionary from high-quality natural images to sparsely represent the source images. Fusion rules are applied to sparse coefficients, and the fused image is reconstructed using a learned dictionary based on these coefficients [44]. This methodology is suitable for merging images obtained from multiple sensors. Compared to conventional multi-scale transformations, our approach produces a dictionary that is more stable and meaningful, as inferred from the training images [45], [46].

Although conventional approaches to IVIF are characterized by their simplicity in structure and ability to provide clear explanations, they often rely on predetermined algorithms or mathematical models. This restriction can impede the effective extraction and integration of critical information from both the visual modalities. Consequently, artifacts or distortions may arise in the fused image, particularly in situations characterized by significant contrast or brightness changes.

B. DL BASED METHODS

Deep Learning (DL)-based techniques have significantly advanced the field of image fusion by leveraging the exceptional capabilities of deep neural networks for robust feature extraction and representation. Initially, DL-based image fusion methodologies utilized pre-trained networks to derive features from source images [47], [48]. For instance, Li [47] decomposed source images into a base layer and detail layer, generating weights based on the features extracted by deep neural networks. These weights were then applied to the detail layer, and a simple averaging method was employed for base-layer fusion. Li also introduced a fusion approach in another study [49] and presented a fusion strategy that involved the utilization of nested connections to mitigate the loss of crucial information. By merging features across different levels, this approach harnessed multi-scale information, ensuring the preservation of both spatial and textural details.

Recognizing the potential of GAN in image fusion, Ma et al. first introduced GAN in IVIF [30]. However, their initial model, which relied exclusively on a discriminator for visible images, resulted in substantial loss of thermal infrared information in the fused output. Addressing this limitation, subsequent research in [33] focused on incorporating dual discriminators into a GAN, which can simultaneously constrain the fused images produced by the generator. Xu et al. proposed a unified image fusion framework [34] that investigated the intrinsic associations between different image fusion tasks, including infrared and visible image fusion. However, this method inadequately retains the thermal radiation information of an infrared image, resulting in blurred edges in the fused image. To overcome the hurdles in GAN training, Ma et al. proposed a novel fusion framework [50], which more fairly integrates visible and infrared images fairly. This method produces a fused image with an enhanced contrast and detailed textural information. Although GAN-based strategies are well suited for the unsupervised task of fusing infrared and visible images, they present substantial training challenges.

The application of auto-encoders (AE) for image fusion is now widely recognized as an influential technique. AE is a type of unsupervised neural network that is capable of efficiently learning data representations by employing encoding and decoding operations. DenseFuse [51] is a remarkable AE-based IVIF method. It leverages MS-COCO [52] for autoencoder pretraining and adopts unique fusion strategies, including addition and L_1-norm for feature fusion. Additionally, Xu et al. [53] introduced a dual-branch AE network that effectively captured both shallow and deep information. Images are decomposed based on infrared and visible differences with appropriate fusion techniques applied during reconstruction. To further optimize the performance of IVIF networks based on AE, some researchers have incorporated attention mechanisms. These mechanisms typically encompass both channel and spatial attention branches. Thus, the network can focus on the salient target areas of the source images, thereby enhancing its feature representational capability. For example, FusionGRAM [54] combines dense connection networks [55] with attention mechanisms to enhance the network's feature extraction capability and adaptively allocate more weight to salient regions. Xu et al. [56] introduced a network that combined skip connections and

TABLE 1. Network configurations. Layer and Size denote the network module and the size of convolutional kernel, respectively.

Module	Layer	Size	Stride	Input Channel	Output Channel	Padding	Activation
PSCNet	SA	-	-	384	384	-	-
	CA	-	-	384	384	-	-
Encoder	MKCBlock	-	-	-	-	1	-
	PSCNet	-	-	-	-	1	-
Decoder	Conv3	3	1	768	384	1	ReLU
	Conv3	3	1	384	192	1	ReLU
	Conv3	3	1	192	96	1	ReLU
	Conv3	3	1	96	48	1	ReLU
	Conv3	3	1	48	1	1	Tan
MKCBlock	Conv3	3	1	16,96,192	32,64,128	1	ReLU
	Conv5	5	1	16,96,192	32,64,128	1	ReLU
	Conv7	7	1	16,96,192	32,64,128	1	ReLU
SA	Conv3	3	1	16,96,192	32,64,128	1	ReLU
	Maxpooling	-	1	16,96,192	32,64,128	1	ReLU
	Averagepooling	-	1	16,96,192	32,64,128	1	GELU
CA	Conv3	3	1	384	384	1	ReLU
	Maxpooling	-	-	384	1	1	ReLU
	Averagepooling	-	-	384	1	1	ReLU

attention mechanisms. The attention module uses three networks with various kernels to enhance multi-scale feature fusion and detail retention. Although attention-based IVIF networks yield positive results, the sequential nature of the spatial and channel attention networks risks losing crucial information during fusion.

To address the aforementioned challenges, we propose AMFusionNet, a network architecture that integrates MKCBlock and PSCNet modules. MKCBlock efficiently extracts diverse features from images using a multi-kernel convolution. PSCNet, which employs parallel spatial and channel attention mechanisms, enhances the representation of essential information, thereby generating high-quality fused images. Furthermore, we incorporated MS-SSIM loss into the loss function to ensure more detailed preservation of the fused images.

III. THE PROPOSED METHOD

As depicted in Fig. 2, AMFusionNet is composed of two main segments: an encoder and decoder. The encoder includes a basic convolutional layer, three MKCBlocks, and a PSCNet, whereas the decoder consists of five convolutional layers. The Tab. 1 lists the network configuration. The encoder is composed of a basic convolutional layer, three MKCBlocks, and a PSCNet. The decoder functions as a component for reconstructing the features and consists of five convolutional layers with kernel sizes of 3×3. The activation functions utilized in the first four layers consist of Batch Normalization (BN) and Parameterized ReLU (PReLU). The fifth layer uses the hyperbolic tangent (Tan) as its activation function.

A. MKCBLOCK

MKCBlock forms an integral part of the encoder comprising convolutional kernels of various sizes (3×3, 5×5, 7×7). The presence of diverse convolutional kernels extends the receptive field of the network by enabling the encoder to extract a rich set of feature information. The MKCBlock architectural diagram is shown in Fig. 3-A.

B. PSCNET

The architectural diagrams of PSCNet and CBAM [57] are shown in Fig. 4 (a) and Fig. 4 (b), respectively. The channel attention network of PSCNet is inspired by the findings presented in CBAM [57]. As depicted in Fig. 4, PSCNet is organized with a spatial attention module and channel attention module. The channel attention module is identical to that found in the CBAM. In the spatial attention module, the inputs to the maxpool and average pool are the intermediate feature maps and original image, respectively. The activation function used in the spatial attention module is GELU. PSCNet exhibits three distinct differentiations from CBAM [57]. First, the PSCNet model incorporates a parallel attention mechanism. Within the field of image fusion, this particular framework enables the integration of features from the channel and spatial attention branches in a channel-wise manner. Consequently, this helps alleviate the loss of information to a certain degree. In addition, the sigmoid function is substituted by the GELU function within the spatial attention module. The GELU activation function possesses nonlinear properties that enable it to preserve the texture information and mitigate gradient vanishing. Finally, the original image is fed into the spatial attention module to minimize information loss.

C. LOSS FUNCTION

The design of the loss function is essential for effective reconstruction of the input images. Therefore, many types of loss functions have been developed in the literature by integrating pixel-based loss L_{pixel} and SSIM loss L_{ssim} [28], [58]. These functions can be mathematically expressed as follows:

$$L_{pixel} = \frac{1}{BCHW}\|O - I\|_2 \tag{1}$$

$$L_{ssim} = 1 - SSIM(O - I) \tag{2}$$

$$SSIM(O, I) = \frac{(2\mu_O\mu_I + C_1)(\sigma_{OI} + C_2)}{(\mu_O^2 + \mu_I^2 + C_1)(\sigma_O^2 + \sigma_I^2 + C_2)} \tag{3}$$

where O and I are the output and input images, respectively; $\| \bullet \|_2$ is the L_2-norm; B represents the batch size; C is the number of image channels; H and W are the height and width, respectively. Notably, channel C of the input and output images are the same; similarly, H and W of the output and input images are also the same. μ_O and μ_I represent the mean pixel values of O and I, respectively; σ_O and σ_I denote the standard deviations of the input and output images, respectively; σ_{OI} represents the covariance between the two images; and C_1 and C_2 are small positive constants.

Because L_{pixel} minimizes the Euclidean distance between the pixel values in input image I and output image O, it does not account for the semantic relationships between the pixels. Moreover, they are susceptible to challenges related to detail losses and excessive smoothing. To address these limitations, we introduce a novel loss function denoted as $L_{MS-SSIM}^{l_1}$, that integrates $L_{MS-SSIM}$ and L_1-norm. These two loss functions enable the fused image to retain the detail and luminance from the input images while enhancing the contrast of the fused

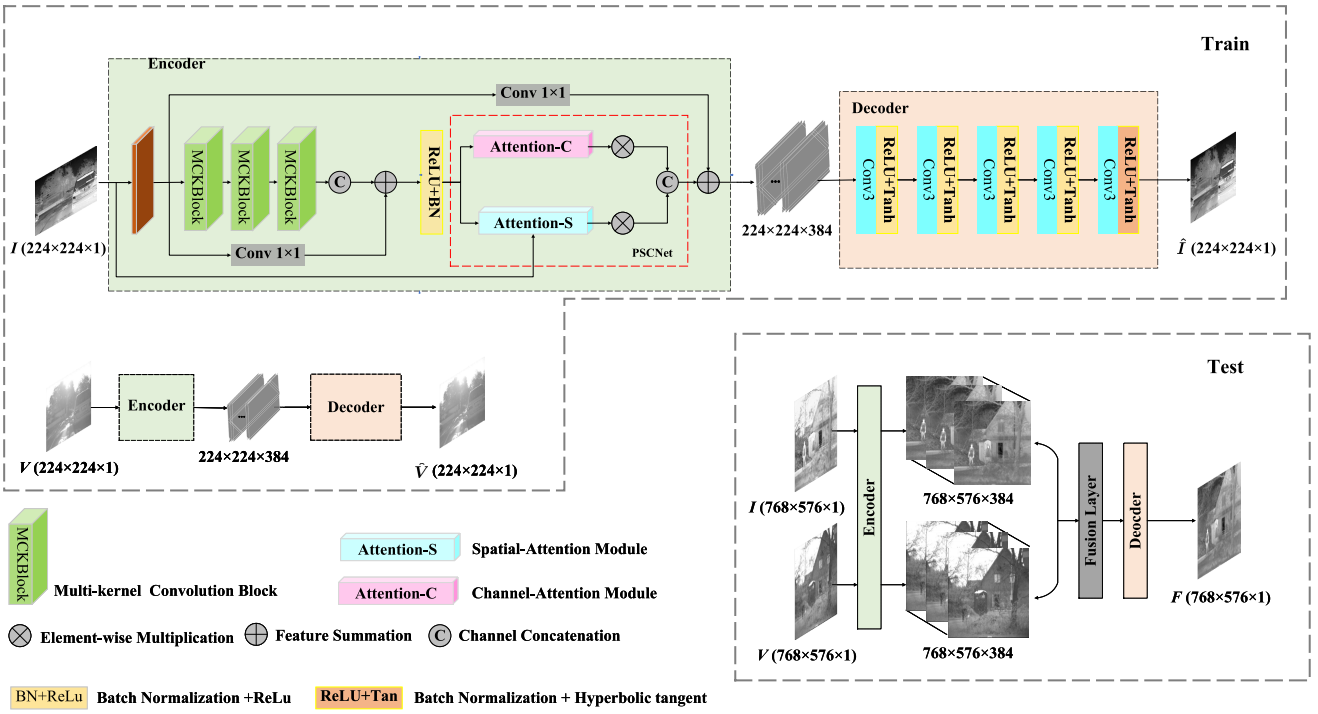

FIGURE 2. The architecture diagram of the AMFusionNet. AMFusionNet is composed of two core network modules, namely PSCNet and MKCBlock.

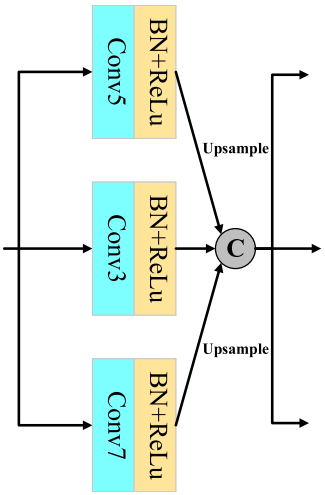

FIGURE 3. The architecture diagram of the MKCBlock.

image. The $L_{MS-SSIM}^{l_1}$ is computed as follows:

$$L_{MS-SSIM}^{l_1} = (1-\beta)L_{MS-SSIM} + \beta L_{l_1} \tag{4}$$

where

$$L_{MS-SSIM} = 1 - l_j(O,I) \cdot \prod_{j=1}^{M} c_j(O,I) s_j(O,I) \tag{5}$$

$$L_{l_1} = \frac{1}{HW} \|O - I\|_1 \tag{6}$$

$$l_j(O,I) = \frac{2\mu_{O_j}\mu_{I_j} + C_1}{\mu_{O_j}^2 + \mu_{I_j}^2 + C_1}, \quad j = \{1, 2, \cdots, M\} \tag{7}$$

$$c_j(O,I) = \frac{2\sigma_{O_j}\sigma_{I_j} + C_2}{\sigma_{O_j}^2 + \sigma_{I_j}^2 + C_2}, \quad j = \{1, 2, \cdots, M\} \tag{8}$$

$$s_j(O,I) = \frac{\sigma_{O_j I_j} + C_3}{\sigma_{O_j}\sigma_{I_j} + C_3}, \quad j = \{1, 2, \cdots, M\} \tag{9}$$

where β is a weight parameter. l_j, $c_j(O,I)$ and $s_j(O,I)$ are the luminance comparison at the j-th scale, the contrast comparison at the j-th scale and the structure comparison at j-th scale, respectively. μ_{O_j} and μI_j are average luminance of O and I at j-th sacle.

Inspired by [59] and [60], the gradient loss function can force fused images to obtain richer texture information, which is defined as

$$L_{grad} = \|\nabla O - \nabla I\|_1 \tag{10}$$

By combining Eq. (1), Eq. (2), Eq. (4), and Eq. (9), the total loss function can be expressed as:

$$L_{total} = \alpha_1 L_{pixel} + \alpha_2 L_{ssim} + \alpha_3 L_{MS-SSIM}^{l_1} + \alpha_4 L_{grad} \tag{11}$$

where α_1, α_2, α_3, and α_4 are positive tuning parameters.

D. FUSION STRATEGY

The fusion layer is essential for generating fused images, and we implement three distinct fusion strategies: (1) the weighted average method, (2) the L_1-norm method, and (3) the mean operator method.

After passing the input images through the encoder, the feature maps were obtained. Fused maps are created using a special fusion strategy to perform the weighted fusion of feature maps from different modalities. Subsequently, the

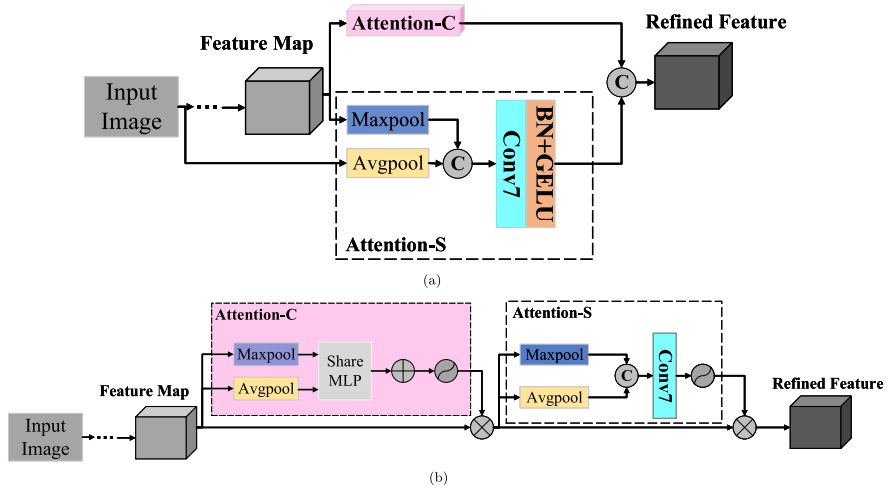

FIGURE 4. (a) Architecture of PSCNet. (b) Architecture of CBAM.

fused maps are input into the decoder to generate a high-resolution fused image.

1) WEIGHTED-AVERAGE METHOD

The weighted average method [61], recognized as a straightforward fusion strategy, has been widely applied in the field of image fusion. This is not only due to its simple design but also its commendable fusion performance. Its formula can be expressed as:

$$f_{fused}^{m}(x,y) = \sum_{i=1}^{2} \frac{f_i^m(x,y)}{2} \tag{12}$$

where $f_i^m(x,y)$ denotes the feature map obtained by the encoder from input images, $f_{fused}^{m}(x,y)$ denotes the fused featured maps, $m \in \{1,2,\cdots,M\}$, M is total channel number of feature maps.

2) THE L_1-NORM METHOD

Inspired by [62], we used a L_1-norm operator to the measure activity. Thus, we obtained the weighted maps ω_i, which were computed as:

$$\omega_i(x,y) = \frac{\sum_{j=1}^{m} \|f_i^j(x,y)\|_1}{\sum_{i=1}^{k}\sum_{j=1}^{m} \|f_i^j(x,y)\|_1} \tag{13}$$

Finally, the fused maps can be obtained as following:

$$f_{fused}^{1:M}(x,y) = \sum_{i=1}^{k} \omega_i(x,y) \times f_i^{1:M}(x,y) \tag{14}$$

3) THE MEAN FILTER OPERATOR METHOD

The mean filter operator is used to process the feature map extracted by the encoder. The weights of the feature maps can be calculated by

$$\omega_i(x,y) = \frac{\varphi(\|f_i(x,y\|_1)}{\sum_{i=1}^{k} \varphi(\|f_i(x,y)\|_1)} \tag{15}$$

where $\varphi(\cdot)$ is a 3×3 mean filter. Consequently, the fused feature maps can be expressed as:

$$f_{fused}^{1:M}(x,y) = \sum_{i=1}^{k} w_i(x,y) \times f_i^{1:M}(x,y) \tag{16}$$

IV. EXPERIMENTS AND DISCUSSIONS

In this section, we describe the experimental settings and environment. Subsequently, we compared the performance of our proposed model with other SOTA models, including one traditional methods, namely MDLatLRR [13], and eight DL-based methods, namely CUFD [53], FusionGAN [30], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], AEFusion [64], and MUFusion [65]. All experiments were conducted using PyTorch on a work station equipped with an Intel Xeon CPU@2.2GHz and an RTX 3090 GPU.

A. DATASETS AND PREPARATION

During the training stage, 180 pairs of images were randomly selected from the dataset [66] to train the proposed model. Before training, all images were converted to grayscale and a center crop of 224×224 pixels was applied to the input images. The pixel intensities of the input images were normalized to the range of -1 to 1. For the training phase, the loss function was optimized using the Adam solver. In the testing phase, we used two datasets, namely TNO [67] and FLIR [68], to evaluate the efficiency and performance of the proposed model.

B. TRAINING SETTINGS AND FUSION METRICS

The training parameters were established using batch image sizes of 12 and 160 training iterations. The hyperparameters within the loss function were empirically determined as $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 2$, $\alpha_4 = 0.005$, $\alpha_5 = 10$, and $\beta = 0.0025$.

A visual assessment of fusion performance can be intricate; hence, we utilized quantitative fusion metrics for an unbiased evaluation. In our study, we adopted eight metrics: entropy (EN) [69], which denotes the informational content of the image; mutual information (MI) [70], which gauges the similarity between input image pairs and the fused output; SSIM [71], which determines the structural resemblance between the source and fused images; average gradient (AG) [72], which is indicative of the image's detail representation and its clarity; standard deviation (SD) [73], which portrays the image's distribution and contrast; spatial frequency (SF) [74], which evaluates the gradient distribution, as well as the image's detail and texture; Q_{abf} [75], which represents the quality of the visual data; and visual information fidelity (VIF) [76], which assesses visual data fidelity; SCD [77] calculates the image quality metric value based on the sum of the correlations of differences. For these metrics, higher values indicate superior performance.

C. EXPERIMENTS ON FUSION STRATEGY

In this section, we evaluate the performance of the three fusion strategies using 40 randomly selected image pairs from the TNO and FLIR datasets. The assessment relies on the average metrics derived from these image pairs. Tab. 2 shows the performance outcomes based on the nine evaluation metrics. Notably, the L_1-norm fusion strategy achieved the four best values for the FLIR dataset and three best values for the TNO dataset. Therefore, subsequent experiments adopted L_1-norm fusion strategy.

TABLE 2. Results of the validation set for choosing the addition strategy. Bold indicates best result.

Dataset: FLIR Dataset			
Method	Summation	Average	L_1-norm
EN	7.3829	**7.3934**	7.3869
AG	**5.9153**	5.7364	5.8207
MI	2.7693	2.8365	**2.8520**
SD	51.3609	**51.8447**	51.5468
SF	**14.9921**	14.6948	14.8068
Q_{abf}	0.4661	0.4598	**0.4732**
SSIM	0.9680	0.9703	**0.9745**
VIF	0.5782	0.5874	**0.5906**
SCD	**1.6975**	1.6924	1.6892
Dataset: TNO Dataset			
Method	Summation	Average	L_1-norm
EN	7.3431	**7.3762**	7.3654
AG	**5.9753**	5.8069	5.8853
MI	2.2406	**2.3294**	2.3226
SD	47.0399	**48.5152**	47.7786
SF	**14.8042**	14.5458	14.6055
Q_{abf}	0.3718	**0.3872**	0.3810
SSIM	0.8396	0.8494	**0.8538**
VIF	0.6535	0.6590	**0.6683**
SCD	1.6922	1.6871	**1.6987**

D. EXPERIMENTAL RESULT AND ANALYSIS

In this section, we evaluate the performance of our proposed model in comparison with other state-of-the-art methods.

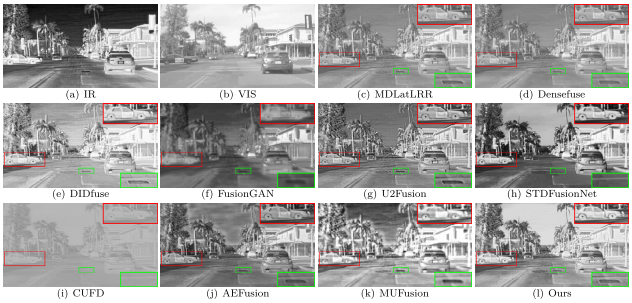

FIGURE 5. Qualitative comparison of the AMFusionNet with seven state-of-the-art methods on *FLIR_04424*. For a clear comparison, we selected a salient region (*i.e.*, the red box and green color box) in each image and zoomed in on it using the same color box, respectively. From top to bottom (from left to right): infrared and infrared image pair, fusion results of the CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65], and AMFusionNet.

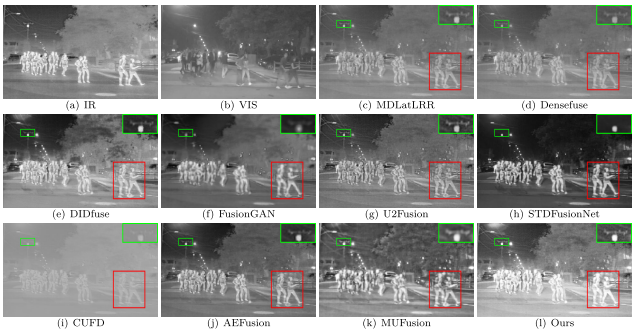

FIGURE 6. Qualitative comparison of the AMFusionNet with seven state-of-the-art methods on *FLIR_07620*. For a clear comparison, we selected a salient region (i.e., the red box and green color box) in each image and zoomed in on it using the same color box, respectively. From top to bottom (from left to right): infrared and infrared image pair, fusion results of the CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65], and AMFusionNet.

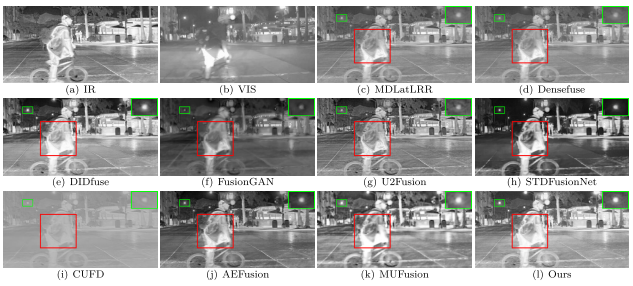

FIGURE 7. Qualitative comparison of the AMFusionNet with seven state-of-the-art methods on *FLIR_09488*. For a clear comparison, we selected a salient region (i.e., the red box and green color box) in each image and zoomed in on it using the same color box, respectively. From top to bottom (from left to right): infrared and infrared image pair, fusion results of the CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65], and AMFusionNet.

We conducted both qualitative and quantitative evaluations of publicly available TNO and FLIR datasets. Qualitative comparison results for the FLIR dataset are shown in Fig. 5-7.

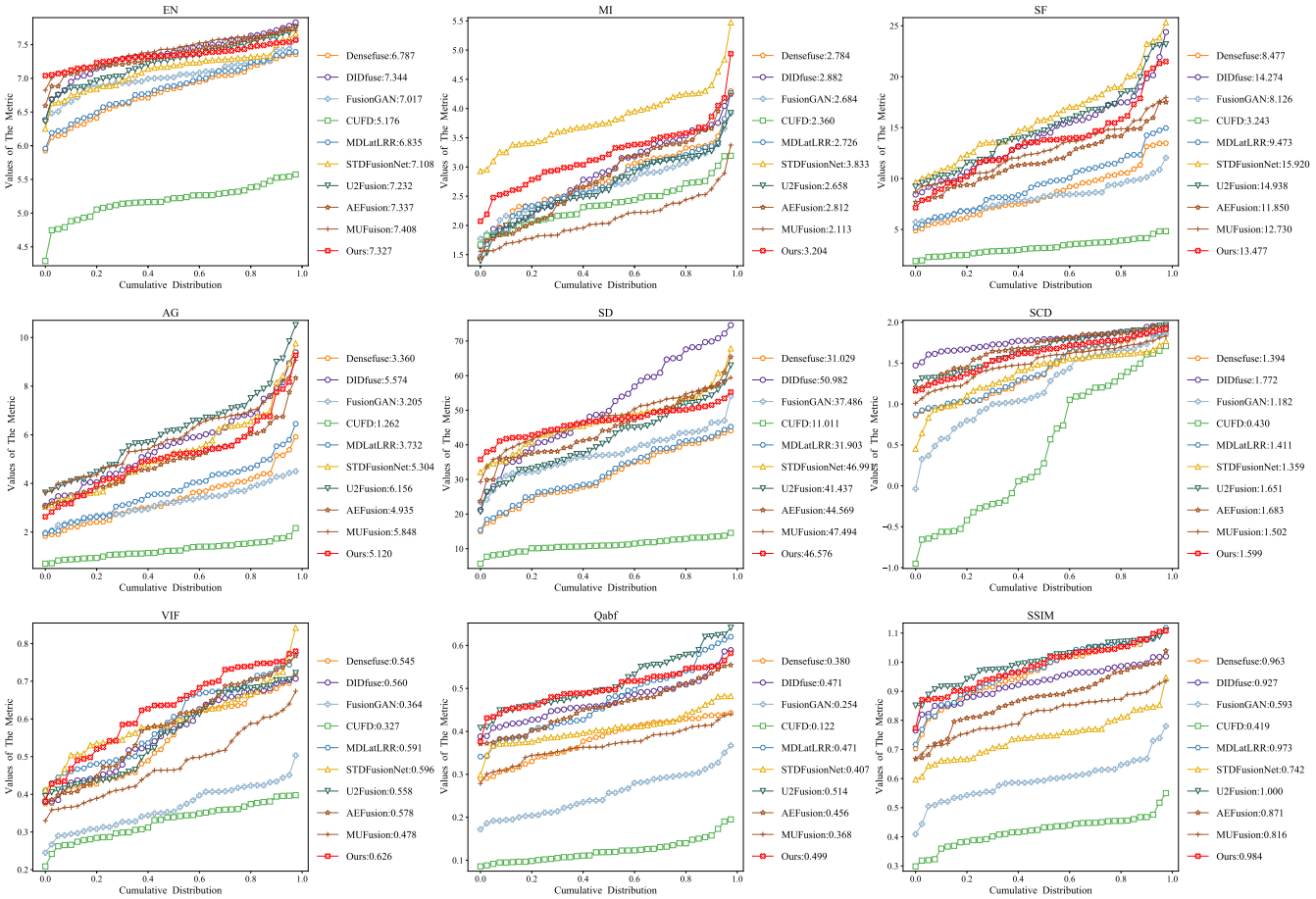

FIGURE 8. Quantitative comparisons of the nine metrics, i.e., EN, MI, SF, AG, SD, SCD, Q_{abf}, VIF and SSIM, on forty image pairs from the FLIR dataset. The nine state-of-the-art methods such as CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65] are used for comparison. A point (x, y) on the curve denotes that there are $(100 * x)$ % percent of image pairs which have metric values no more than y. Our proposed AMFusionNet is indicated by a red X mark line.

1) QUALITATIVE RESULTS

To directly compare the effectiveness of our proposed algorithm for image fusion with existing methods, we chose three representative pairs of source images from the FLIR dataset. The fusion results of the different algorithms are presented in Fig. 5-7. In Fig. 5-7, we identified two regions for comparative analysis: a background region and a salient target region. As shown in Fig. 5, CUFD, FusionGAN and AEFusion overlooked the essential thermal emission details, thereby missing the thermal information of the car. Although STDFusionNet, MUFusion and U2Fusion highlight the salient target, they fall to capture key background elements such as the manhole cover. In contrast, AMFusionNet effectively differentiates the target from its background, while preserving the texture of the visible image. As shown in Fig. 6, each algorithm consistently captured the thermal details, thereby enabling a distinct separation between the target and its background. However, our method distinguishes itself by preserving intricate background details. It effectively captured the thermal details of the infrared image and the texture of the visible image, as shown in the magnified red box. This exceptional performance serves as an evidence of the effectiveness of the proposed approach.

From Fig. 7, it appears that both FusionGAN Densefuse and MDLatLRR potentially compromise the thermal radiation details of the infrared salient target, which is attributable to the constraints of the visible light image under low-light conditions. Although STDFusionNet, AEFusion, and MUFusion captured most of the infrared data, their preservation of background details was lacking, as indicated by the blurred boundaries of the tree. U2Fusion retains details from both the infrared target and background texture. CUFD [53] loss of both infrared information and fine details, indicating a significant limitation in its ability to preserve key image attributes. Additionally, although DIDfuse effectively retains thermal information from infrared images and textural details, it tends to produce background artifacts.

In comparison, the proposed algorithm clearly stands out, delivering fused images with enhanced target brightness, sharper edge contours, and superior retention of intricate detail. Noteworthy examples include the manhole cover in Fig. 5 and cyclist in Fig. 7.

2) QUANTITATIVE RESULTS

We evaluated the performance across 40 image pairs from the FLIR dataset using nine established quantitative metrics, as illustrated in Fig. 8 and Tab. 3. Compared to nine other state-of-the-art algorithms, our algorithm secured the leading position in the SD evaluation metric. In addition, it exhibited exceptional performance in the EN, SCD, VIF, SSIM, and SF which were comparable to those of the top performers in the other methods. Even with second-place standing, our algorithm consistently exhibited competitive performance, highlighting its advantages in producing high-quality fused images.

As illustrated in Fig. 5-7 and detailed in Tab. 3, the efficacy of the algorithm can be assessed subjectively through visual inspection and objectively by using nine distinct metrics. Acknowledging the unique strengths of each algorithm, as reflected in these metrics, we introduce an evaluation metric called the normalized evaluation index, denoted as

$$\xi = \sum_{j=1}^{N=9} \frac{M_j(F, V, I))}{\max M_j(F, V, I)} \tag{17}$$

where $M_j(\bullet)$ indicates the formula of the evaluation metric; $N = 9$ denotes the nine evaluation metrics; F, V and I represent the fused, visible, and infrared images, respectively.

Based on the aggregated normalized values from these nine metrics, the proposed algorithm has the highest value. This performance distinction further emphasizes the superiority of the proposed algorithm.

E. GENERALIZATION ANALYSIS

Evaluating the generalization capability of a DL model is crucial to determine its overall efficacy. To assess the generalization performance of our AMFusionNet model, we tested it on image pairs from the TNO dataset as the model was trained on the FLIR dataset.

1) QUALITATIVE RESULTS

Examining Fig. 9-11, we can observe the fusion results of the various methods. Our algorithm effectively retains the thermal information from the infrared images and accentuates the boundaries of prominent targets in the fused images. Compared to other methods, our technique preserves more background details, offers heightened contrast, clarifies finer details, and renders targets more distinguishable. For instance, consider Fig. 9, where the fusion of MDLatLRR fails to account for the radiation details of the infrared target. Although DIDfuse, AEFusion, and STDFusionNet effectively identify significant infrared targets in their fusion results, their outputs show smoothing effects, leading to a loss of thermal detail, such as the subdued thermal radiation details of streetlights. In Fig. 10, in addition to the fusion images of MDLatLRR, U2Fusion, and AEFusion, the fusion results of almost all methods successfully retained the infrared information. However, a drawback was observed in the fusion results of Densefuse, CUFD, and FusionGAN,

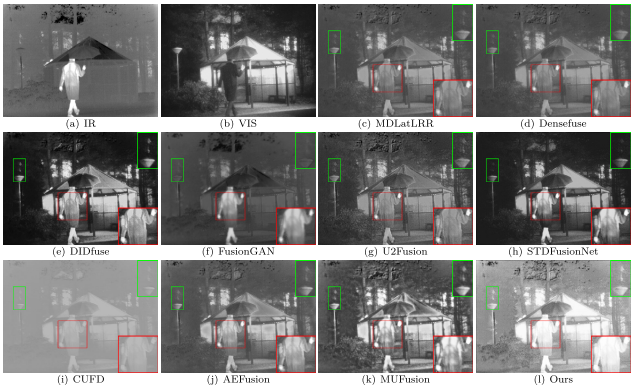

FIGURE 9. Qualitative comparison of the AMFusionNet with seven state-of-the-art methods on *soldier_behind_smoke*. For a clear comparison, we selected a salient region (*i.e.*, the red box and green color box) in each image and zoomed in on it using the same color box, respectively. From top to bottom (from left to right): visible and infrared image pair, fusion results of the CUFD [53], MDLatLRR [13], STDFusionNet [78], FusionGAN [30], Densefuse [51], DIDfuse [63], U2Fusion [34], and our AMFusionNet.

in which the background information was excessively smoothed. In addition, in the fusion results of MUFusion, the contrast between the foreground and background information is not pronounced, which makes it difficult to distinguish the significant target. In Fig. 11, the fusion results of each method easily distinguish significant targets from the background; however, each method has drawbacks. For instance, STDFusionNet loses cloud information, and MDLatLRR, CUFD, and Densefuse fail to capture information about people, as shown in the green box. The MUFusion fusion method retains detailed information from the visible image and thermal radiation information from the infrared and generates numerous artifacts, such as cloud information that does not exist in the source image. By contrast, our algorithm produces fusion results that clearly distinguish infrared targets with sharp boundaries.

2) QUANTITATIVE RESULTS

To provide an objective evaluation of the performance across different fusion algorithms, we analyzed 40 image pairs from the TNO dataset. Fig. 12 and Tab. 4 presents the fusion outcomes, clearly demonstrating the robust performance of our algorithm across the nine metrics. On the TNO dataset, our method clinched the top scores in EN, AG, MI, SD, SF, SCD and Normalized Values and secured a commendable second place in VIF. A leading EN score indicates that our fused images encompass more information richness than other methods. High AG, SD, and SF scores highlight information-rich and visually striking results that adeptly preserve details, contrast, and texture from the source images. Leading scores in MI and SCD further attest to the capability of our technique to optimally transfer information from the source images to the fused outputs. Furthermore, AMFusionNet achieved the best outcome in terms of Normalized Values, suggesting that the proposed approach exhibits an

TABLE 3. Quantitative comparisons of the nine metrics, i.e., EN, AG, MI, SD, SF, Q_{abf}, SSIM, VIF, SCD, on the 40 image pairs from the FLIR dataset. **Bold** indicates the best result and _ indicates the second best result.

Methods	EN	AG	MI	SD	SF	Q_{abf}	SSIM	VIF	SCD	Normalized Values
CUFD	5.1761	1.2616	2.3602	11.0114	3.243	0.1216	0.4195	0.3271	0.4303	3.4571
MDLatLRR	6.8351	3.7318	2.7263	31.9031	9.473	0.4707	0.9672	0.5906	1.411	7.1220
STDFusionNet	7.1080	5.3042	**3.8334**	46.9905	**15.9198**	0.4066	0.7367	**0.596**	1.3588	8.0235
U2Fusion	7.2324	**6.1577**	2.6577	41.4372	14.9383	**0.5145**	**0.9941**	0.5584	1.6513	8.2733
FusionGAN	7.0167	3.2046	2.6836	37.4861	8.1258	0.2536	0.5951	0.3638	1.1815	5.7669
Densefuse	6.7865	3.3599	2.7839	31.0294	8.4774	0.3801	0.9569	0.5446	1.3938	6.7173
DIDfuse	7.3441	5.5742	2.8822	50.9818	14.2743	0.4713	0.9166	0.5602	**1.7719**	8.3047
AEFusion	7.337	4.935	2.812	44.569	11.850	0.456	0.871	0.578	1.683	7.8090
MUFusion	**7.408**	5.848	2.113	47.494	12.73	0.368	0.816	0.478	1.599	7.4549
Ours	7.3841	5.8207	2.8520	**51.5468**	14.8068	0.4732	0.9745	0.5906	1.6892	**8.4529**

TABLE 4. Quantitative comparisons of the nine metrics, i.e., EN, AG, MI, SD, SF, Q_{abf}, SSIM, VIF, SCD, on the 40 image pairs from the TNO dataset. **Bold** indicates the best result and _ indicates the second best result.

Methods	EN	AG	MI	SD	SF	Q_{abf}	SSIM	VIF	SCD	Normalized Values
CUFD	4.8132	0.8669	2.1764	8.9437	2.1749	0.0999	0.3824	0.2756	0.9824	3.3481
MDLatLRR	6.3904	2.7937	2.1048	25.5972	7.3137	**0.4427**	**1.0158**	0.6195	1.6237	6.8203
STDFusionNet	6.8700	4.2318	**3.2912**	39.734	11.7147	0.4376	0.7982	**0.694**	1.4352	7.8635
U2Fusion	6.9655	4.94	1.1924	36.5903	11.6162	0.4266	0.9588	0.6066	**1.7856**	7.4901
FusionGAN	6.5761	2.41691	2.341	31.1199	6.2466	0.2328	0.6603	0.4201	1.3955	5.6566
Densefuse	6.3518	2.5148	2.216	24.7829	6.3794	0.3506	1.0127	0.5727	1.6056	6.4318
DIDfuse	7.0061	4.2942	2.3468	46.8854	11.2839	0.4027	0.8658	0.6235	1.7837	7.8071
AEFusion	7.027	3.463	2.357	38.505	7.760	0.358	0.790	0.554	1.693	6.9142
MUFusion	7.219	4.614	1.940	45.244	9.705	0.365	0.799	0.540	1.569	7.2328
Ours	**7.3654**	**5.8853**	2.3226	**47.7786**	**14.6055**	0.3810	0.8538	0.6683	1.6987	**8.3211**

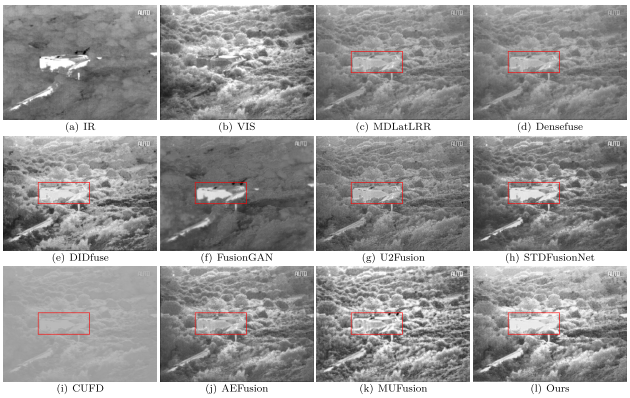

FIGURE 10. Qualitative comparison of the AMFusionNet with seven state-of-the-art methods on *Kaptein_1654*. For a clear comparison, we selected a salient region (*i.e.*, the red box and green color box) in each image and zoom in on it using the same color box, respectively. From top to bottom (from left to right): visible and infrared image pair, fusion results of the CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65], and AMFusionNet.

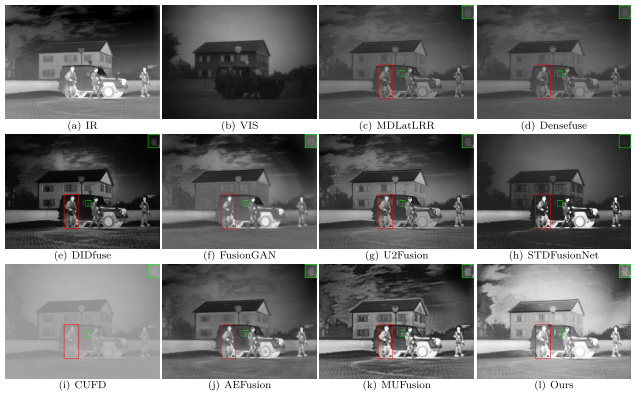

FIGURE 11. Qualitative comparison of the AMFusionNet with seven state-of-the-art methods on *Jeep*. For a clear comparison, we selected a salient region (*i.e.*, the red box and green box) in each image and zoomed in it use same color box, respectively. From top to bottom (from left to right): visible and infrared image pair, fusion results of the CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65], and AMFusionNet.

optimal overall performance. Specifically, it achieves a better balance between preserving the thermal radiation information in infrared images and capturing texture details in visible images.

In conclusion, the images produced by our approach are superior for visual perception because they offer improved contrast, clear salient targets, and minimal texture degradation. Objectively and subjectively, these attributes underscore the robust generalization capabilities of our method. According to Eq. (17), we calculated the normalized evaluation metrics for the fused images generated by AMFusionNet and the other seven algorithms, as listed in Tab. 4.

F. ABLATION EXPERIMENT

In this section, we evaluate the impact of the attention mechanisms and MS-SSIM loss function on the performance of the AMFusionNet. We explored three variations: AMFusionNet without the attention mechanism (AMFusionNet-SC), AMFusionNet employing a parallel form of the attention mechanism (PSCNet), and AMFusionNet excluding the MS-SSIM loss function (AMFusionNet-MSSSIM).

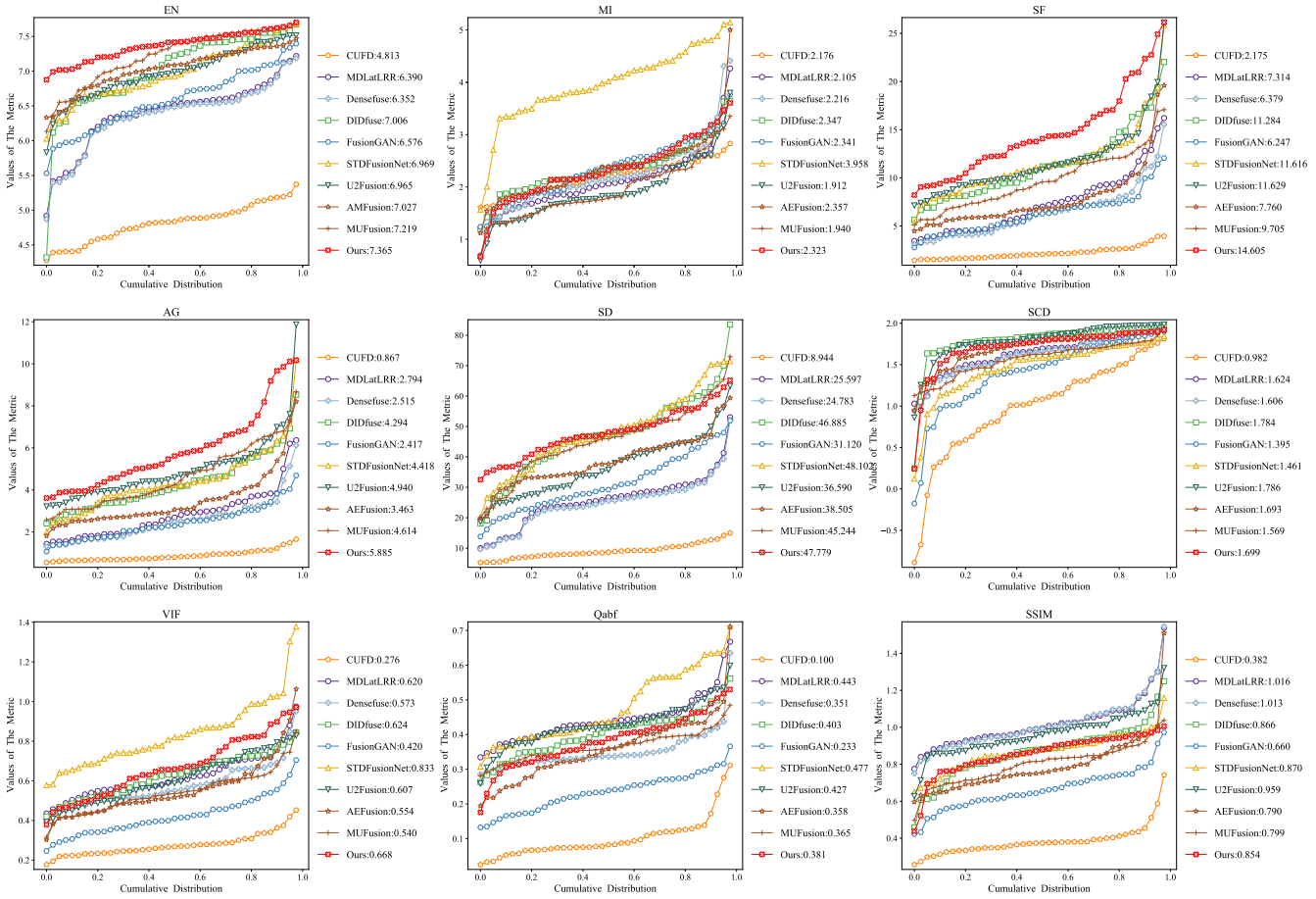

FIGURE 12. Quantitative comparisons of the nine metrics, i.e., EN, MI, SF, AG, SD, SCD, Q_{abf}, VIF and SSIM, on forty image pairs from the FLIR dataset. The nine state-of-the-art methods such as CUFD [53], MDLatLRR [13], Densefuse [51], DIDfuse [63], FusionGAN [30], U2Fusion [34], STDFusionNet [78], AEFusion [64], MUFusion [65] are used for comparison. A point (x, y) on the curve denotes that there are $(100 * x)$ % percent of image pairs which have metric values no more than y. Our proposed AMFusionNet is indicated by a red X mark line.

1) ATTENTION MECHANISM IMPACT

The role of the attention mechanism is crucial for tasks related to feature extraction and fusion [79], [80]. To assess its importance, we performed an experiment that excluded the attention module while maintaining the remaining network components.

Fig. 13 shows the fusion results of both AMFusionNet and AMFusionNet-SC. Notably, AMFusionNet emphasizes distinct targets in salient areas and preserves the intricate textures of background regions. In contrast, AMFusionNet-SC struggles to retain key infrared image details and delineate sharp boundaries for salient targets, unlike AMFusionNet. Quantitative metrics further highlight the enhanced performance of AMFusionNet compared with that of AMFusionNet-SC. Tab. 5 shows that AMFusionNet achieved the highest scores in five metrics and a commendable second place in one metric.

2) ANALYSIS OF MS-SSIM LOSS

Integrating the MS-SSIM loss function bolsters the network's ability to preserve the texture details. This enhancement arises from the technique of comparing features between the input and output images at various scales, making the network more sensitive to subtle image changes. Consequently, the network is less prone to excessive image smoothing, thereby ensuring the preservation of finer image details. Fig. 13 presents a qualitative comparison between AMFusionNet and AMFusionNet-MSSSIM using the image sets *men_in_front_of_house* and *Kaptein_19*. Comparing the fusion outcomes, AMFusionNet was observed to preserve more texture details, as exemplified by the tree textures, compared with AMFusionNet-MSSSIM. Tab. 5 indicates AMFusionNet's superior performance over AMFusionNet-MSSSIM in four key metrics: EN, AG, SD, and SF. EN measures the amount of information contained in a fused image, AG measures the richness of the edge and texture information in the image, SF measures the complexity of the texture, and SD measures image contrast. All three indicators intuitively indicate that AMFusionNet performed better than AMFusionNet-MSSSIM.

3) IMPACT OF THE PARALLEL ATTENTION MECHANISM

Building on our earlier discussion, the sequential attention mechanism inherently suffers from information loss given its

TABLE 5. Quantitative comparisons of the four metrics, *i.e.*, EN, AG, MI, SD, SF, Q_{abf}, SSIM, VIF, and SCD, on 40 image pairs from the TNO dataset. **Bold** indicates the best result.

	AMFusionNet-SC	AMFusionNet-MSSSIM	PSCNet	AMFusionNet
EN	7.2525	7.225	7.2579	**7.3654**
AG	4.5848	4.8944	4.8083	**5.8853**
MI	2.536	2.239	**3.0903**	2.3226
SD	45.8586	42.0822	44.7442	**47.7786**
SF	11.6601	11.9025	12.2281	**14.6055**
Q_{abf}	0.4253	0.3883	**0.4911**	0.3718
SSIM	0.8892	0.9666	**0.9959**	0.8538
VIF	0.6644	**0.6879**	0.6015	0.6683
SCD	1.7023	**1.7717**	1.5722	1.6987

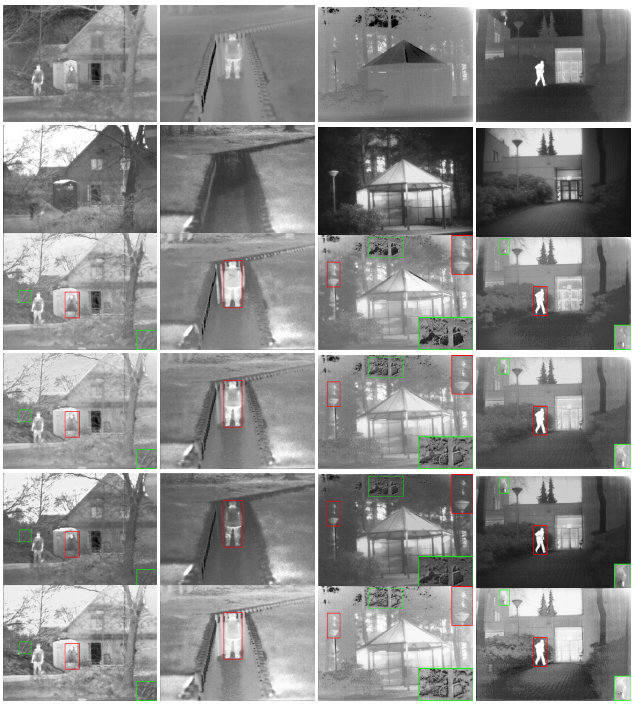

FIGURE 13. A qualitative comparison of AMFusionNet with AMFusion-SC, as well as AMFusion-MSSSIM and PSCNet on *men_in_front_of_house*, was conducted on *Kaptein_1123*, *Kaptein_19*, and *solid_inbranch*. For a clear comparison, we selected a salient region (*i.e.*, the red box and green color box) in each image and zoomed in on it using the same color box, respectively. From top to bottom (from left to right): infrared and visible image pair, fusion results of AMFusionNet-SC, AMFusionNet-MSSSIM, PSCNet, and AMFusionNet.

reliance on the outputs of the preceding steps. This limitation stems from the constrained capacity of the mechanism to effectively merge features from both the channel and the spatial attention branches. From a subjective standpoint, as illustrated in Fig. 13, PSCNet struggles to conserve infrared thermal radiation details, resulting in a fused image with overly smoothed features. Conversely, AMFusionNET effectively retains both forms of information. From an objective perspective, AMFusionNet outperformed PSCNet in four key metrics: EN, AG, SF, and SD. These observations corroborate the results of our theoretical analysis.

V. CONCLUSION

In this paper, we propose a network framework called AMFusioNNET for IVIF, which is based on MKCBlock, PSCNet and MS-SSIM loss. The MCKBlock contains three distinct types of convolutions, each characterized by a unique convolution kernel size. These convolutional operations enable feature extraction at various scales, thereby enhancing the ability of the network to capture complex and detailed feature information. PSCNet, which is designed based on the parallel attention mechanism, allows the network to attend to salient information from the source image. In PSCNet, GELU was also introduced to replace ReLU, which allows the network to retain more detailed information. The introduction of MS-SSIM guided AMFusionNet to compute the similarity between the fused image and original image at multiple scales, which, to a certain extent, mitigated the information loss caused by the depth network. Various experiments have shown that the fusion image obtained by our method achieves competitive results both subjectively and objectively.

REFERENCES

[1] X. Song, X.-J. Wu, and H. Li, "A medical image fusion method based on MDLatLRRv2," 2022, *arXiv:2206.15179*.
[2] Y. Zang, D. Zhou, C. Wang, R. Nie, and Y. Guo, "UFA-FUSE: A novel deep supervised and hybrid model for multifocus image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–17, 2021.
[3] J. Liu, J. Shang, R. Liu, and X. Fan, "Attention-guided global–local adversarial learning for detail-preserving multi-exposure image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5026–5040, Aug. 2022.
[4] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.
[5] 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 (CVPRW)*, Jun. 2020, pp. 468–478.
[6] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, and Y. Qiao, "Pedestrian detection with unsupervised multispectral feature learning using deep neural networks," *Inf. Fusion*, vol. 46, pp. 206–217, Mar. 2019.
[7] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "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. (CVPR)*, Jun. 2022, pp. 5792–5801.
[8] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, vol. 63, pp. 166–187, Nov. 2020.
[9] X. Zeng, J. Long, S. Tian, and G. Xiao, "Random area pixel variation and random area transform for visible-infrared cross-modal pedestrian re-identification," *Expert Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119307.
[10] J. Hou, D. Zhang, W. Wu, J. Ma, and H. Zhou, "A generative adversarial network for infrared and visible image fusion based on semantic segmentation," *Entropy*, vol. 23, no. 3, p. 376, Mar. 2021.
[11] W. Gan, X. Wu, W. Wu, X. Yang, C. Ren, X. He, and K. Liu, "Infrared and visible image fusion with the use of multi-scale edge-preserving decomposition and guided image filter," *Infr. Phys. Technol.*, vol. 72, pp. 37–51, Sep. 2015.

[12] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[13] 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.

[14] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infr. Phys. Technol.*, vol. 76, pp. 52–64, May 2016.

[15] B. Zhang, X. Lu, H. Pei, and Y. Zhao, "A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled shearlet transform," *Infr. Phys. Technol.*, vol. 73, pp. 286–297, Nov. 2015.

[16] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[17] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[18] 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, Sep. 2016.

[19] K. Madheswari and N. Venkateswaran, "Swarm intelligence based optimisation in thermal image fusion using dual tree discrete wavelet transform," *Quant. Infr. Thermography J.*, vol. 14, no. 1, pp. 24–43, Jan. 2017.

[20] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infr. Phys. Technol.*, vol. 82, pp. 8–17, May 2017.

[21] C. Xu, Q. Li, Q. Zhou, X. Jiang, D. Yu, and Y. Zhou, "Asymmetric cross-modal activation network for RGB-T salient object detection," *Knowl.-Based Syst.*, vol. 258, Dec. 2022, Art. no. 110047.

[22] P. Wani, K. Usmani, G. Krishnan, T. O'Connor, and B. Javidi, "Lowlight object recognition by deep learning with passive three-dimensional integral imaging in visible and long wave infrared wavelengths," *Opt. Exp.*, vol. 30, no. 2, p. 1205, 2022.

[23] X. Wang, Z. Yu, S. De Mello, J. Kautz, A. Anandkumar, C. Shen, and J. M. Alvarez, "FreeSOLO: Learning to segment objects without annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14156–14166.

[24] X. Shen, A. A. Efros, A. Joulin, and M. Aubry, "Learning co-segmentation by segment swapping for retrieval and discovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 5078–5088.

[25] Y. Yuan, J. Chu, L. Leng, J. Miao, and B.-G. Kim, "A scale-adaptive object-tracking algorithm with occlusion detection," *EURASIP J. Image Video Process.*, vol. 2020, no. 1, pp. 1–15, Dec. 2020.

[26] Y. Wang, X. Wei, X. Tang, H. Shen, and H. Zhang, "Adaptive fusion CNN features for RGBT object tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7831–7840, Jul. 2022.

[27] 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, Jun. 2022.

[28] Y. Long, H. Jia, Y. Zhong, Y. Jiang, and Y. Jia, "RXDNFuse: A aggregated residual dense network for infrared and visible image fusion," *Inf. Fusion*, vol. 69, pp. 128–141, May 2021.

[29] 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.v.

[30] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[31] 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.

[32] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3954–3960.

[33] 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.

[34] 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.

[35] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Inf. Fusion*, vol. 12, no. 2, pp. 74–84, Apr. 2011.

[36] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.

[37] X. Zhang, Y. Ma, F. Fan, Y. Zhang, and J. Huang, "Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 34, no. 8, p. 1400, 2017.

[38] 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, Jul. 2015.

[39] A. V. Vanmali and V. M. Gadre, "Visible and NIR image fusion using weight-map-guided Laplacian–Gaussian pyramid for improving scene visibility," *Sādhanā*, vol. 42, no. 7, pp. 1063–1082, Jul. 2017.

[40] Z. Ren, G. Ren, and D. Wu, "Fusion of infrared and visible images based on discrete cosine wavelet transform and high pass filter," *Soft Comput.*, vol. 27, no. 18, pp. 13583–13594, Sep. 2023.

[41] 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.

[42] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infr. Phys. Technol.*, vol. 83, pp. 94–102, Jun. 2017.

[43] Y. Li, G. Liu, D. P. Bavirisetti, X. Gu, and X. Zhou, "Infrared-visible image fusion method based on sparse and prior joint saliency detection and LatLRR-FPDE," *Digit. Signal Process.*, vol. 134, Apr. 2023, Art. no. 103910.

[44] Y. Yang, C. Gao, Z. Ming, J. Guo, E. Leopold, J. Cheng, J. Zuo, and M. Zhu, "LatLRR-CNN: An infrared and visible image fusion method combining latent low-rank representation and CNN," *Multimedia Tools Appl.*, vol. 82, no. 23, pp. 36303–36323, Sep. 2023.

[45] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Inf. Fusion*, vol. 40, pp. 57–75, Mar. 2018.

[46] W. Ding, D. Bi, L. He, and Z. Fan, "Infrared and visible image fusion method based on sparse features," *Infr. Phys. Technol.*, vol. 92, pp. 372–380, Aug. 2018.

[47] 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. (ICPR)*, Aug. 2018, pp. 2705–2710.

[48] L. Zhang, H. Li, R. Zhu, and P. Du, "An infrared and visible image fusion algorithm based on ResNet-152," *Multimedia Tools Appl.*, vol. 81, no. 7, pp. 9277–9287, Mar. 2022.

[49] 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.

[50] 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.

[51] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.

[53] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Comput. Vis. Image Understand.*, vol. 218, Apr. 2022, Art. no. 103407.

[54] J. Wang, X. Xi, D. Li, and F. Li, "FusionGRAM: An infrared and visible image fusion framework based on gradient residual and attention mechanism," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[56] D. Xu, N. Zhang, Y. Zhang, Z. Li, Z. Zhao, and Y. Wang, "Multi-scale unsupervised network for infrared and visible image fusion based on joint attention mechanism," *Infr. Phys. Technol.*, vol. 125, Sep. 2022, Art. no. 104242.
[57] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[59] 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, Dec. 2022.
[60] H. Xu, F. Fan, H. Zhang, Z. Le, and J. Huang, "A deep model for multi-focus image fusion based on gradients and connected regions," *IEEE Access*, vol. 8, pp. 26316–26327, 2020.
[61] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4724–4732.
[62] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039.
[63] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," 2020, *arXiv:2003.09210*.
[64] B. Li, J. Lu, Z. Liu, Z. Shao, C. Li, Y. Du, and J. Huang, "AEFusion: A multi-scale fusion network combining axial attention and entropy feature aggregation for infrared and visible images," *Appl. Soft Comput.*, vol. 132, Jan. 2023, Art. no. 109857.
[65] C. Cheng, T. Xu, and X.-J. Wu, "MUFusion: A general unsupervised image fusion network based on memory unit," *Inf. Fusion*, vol. 92, pp. 80–92, Apr. 2023.
[66] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12484–12491.
[67] A. Toet, "Progress in color night vision," *Opt. Eng.*, vol. 51, no. 1, Feb. 2012, Art. no. 010901.
[68] T. Imaging. (Jun. 2023). *FLIR Data Set Dataset*. Accessed: Oct. 17, 2023. [Online]. Available: https://universe.roboflow.com/thermal-imaging-0hwfw/flir-data-set
[69] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.
[70] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, p. 313, 2002.
[71] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[72] J. Wu, H. Huang, Y. Qiu, H. Wu, J. Tian, and J. Liu, "Remote sensing image fusion based on average gradient of wavelet transform," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, Jul. 2005, pp. 1817–1821.
[73] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, p. 355, 1997.
[74] 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.
[75] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, p. 308, 2000.
[76] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013.
[77] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, Dec. 2015.
[78] 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.
[79] H. Wan, J. Chen, Z. Huang, Y. Feng, Z. Zhou, X. Liu, B. Yao, and T. Xu, "Lightweight channel attention and multiscale feature fusion discrimination for remote sensing scene classification," *IEEE Access*, vol. 9, pp. 94586–94600, 2021.
[80] H. Xiao, Q. Liu, and L. Li, "MFMANet: Multi-feature multi-attention network for efficient subtype classification on non-small cell lung cancer CT images," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104768.

QIAN XU received the B.E. degree in measurement and control technology and instrumentation from Shenyang University, China, in 2016, and the M.S. degree in control theory and control engineering from China Jiliang University, China, in 2019. He is currently pursuing the Ph.D. degree with Zhejiang University. His research interests include image processing, pattern recognition, and image fusion.

YUAN ZHENG received the M.Sc. degree from the Illinois Institute of Technology, in 2015, the B.Sc. degree from Guangxi University, in 2016, and the Ph.D. degree from Zhejiang University, in 2021. He is currently a Lecturer with the School of Computer Science, Civil Aviation Flight University of China. His current research interests include path planning of UTM systems and computer vision.