DePF: A Novel Fusion Approach based on Decomposition Pooling for Infrared and Visible Images

Hui Li[1*], Yongbiao Xiao[1], Chunyang Cheng, Zhongwei Shen and Xiaoning Song

***Abstract*—Infrared and visible image fusion is a crucial technique in the field of computer vision, aiming to create synthetic images that simultaneously capture salient features and rich texture details. These fused images play a pivotal role in enhancing various downstream tasks. However, existing fusion methods often encounter challenges such as texture loss and deficiencies in edge information, resulting in less than optimal fusion outcomes. Additionally, straight-forward up-sampling techniques struggle to preserve source information from multi-scale features. To tackle these issues, a novel fusion network based on the decomposition pooling (de-pooling) manner is proposed, termed as DePF. DePF features a de-pooling-based encoder designed to extract multi-scale image and detail features from source images concurrently. Furthermore, a spatial attention model aggregates these salient features, and the decoder employs a de-pooling reversed operation instead of the typical up-sampling operator to reconstruct the fused features. Unlike conventional max-pooling techniques, the de-pooling layer preserves abundant detail information, facilitating a richer texture and multi-scale information retention during the reconstruction phase. Significantly, our approach exhibits remarkable efficiency, requiring merely 23 ms to integrate a pair of infrared and visible images, each with dimensions of 640×480. Furthermore, empirical findings corroborate the exceptional fusion efficacy of our methodology in the domains of object detection and noise-related assessments, surpassing the performance of contemporary techniques within numerous image fusion benchmarks.**

***Index Terms*—image fusion, decomposition pooling, multi-scale features, detail features, deep learning.**

I. Introduction

IMAGE fusion task aims to extract complementary information from source images and generate composite images containing rich information. For infrared and visible images, the infrared modality contains salient thermal radiation information but fewer details, while the visible modality involves rich texture information. Infrared and visible image fusion task is a technique that combines information from images captured in the infrared spectrum and the visible spectrum to create a single composite image with enhanced information and improve visual perception. As an important image processing technology, image fusion has been widely used in areas such as digital remote sensing [1], [2], surveillance [3], [4], agriculture [5], classification [6], object detection [7], [8], [9] and object tracking [10], [11]. Therefore, how to preserve these complementary information from these two modalities is the key issue.

Before the rise of deep learning, most image fusion methods are based on signal processing operations, such as multi-scale transform (MST) [12], [13] based methods, sparse representation (SR) [14], [15], [16], [17] and low-rank representation (LRR) [18], [19], [20] based methods. Non-deep learning methods usually include the following three steps: feature extraction, feature fusion and image reconstruction. Although these image fusion methods can synthesize satisfactory images in certain scenarios, they still contain limitation. On the one hand, manually designed fusion strategies cannot adapt to complex scenes. On the other hand, when extracting features from multi-modal images, the feature differences of each modality are usually not considered, which is difficult to comprehensively capture the features of source images. These problems prevent non-deep learning methods from being extended in practical applications.

With the development of deep learning, many deep learning based fusion methods are proposed to address the above drawbacks and the fusion performance has been improved. Deep learning based image fusion methods can be roughly divided into two categories: auto-encoder based methods[21], [22], [23] and end-to-end fusion networks[24], [25], [26].

In recent years, some typical auto-encoder based networks have been proposed [27], [28]. The auto-encoder architecture is introduced into image fusion mainly because of the insufficient multi-modal dataset, which is difficult to train a complex end-to-end fusion network. With the training strategy of auto-encoder, a large of single modality datasets can be used to train the encoder (feature extraction) and decoder (image reconstruction)[21]. Then, manually designed fusion strategies are applied into auto-encoder framework in testing phase. The auto-encoder based fusion methods are generally composed of an encoder, a fusion strategy block and a decoder. These methods can make full use of the excellent nonlinear fitting ability of neural networks and utilize unsupervised ways to improve the quality of generated images. However, most of them are unable to sufficiently extract features, and the

This work was supported by the National Natural Science Foundation of China (62202205), the National Social Science Foundation of China(21&ZD166), the Natural Science Foundation of Jiangsu Province, China(BK20221535), and the Fundamental Research Funds for the Central Universities (JUSRP123030).

Hui Li, Yongbiao Xiao, Chunyang Cheng and Xiaoning Song are with International Joint Laboratory on Artificial Intelligence of Jiangsu Province, School of Artificial Intelligence and Computer Science, Jiangnan University, 214122, Wuxi, China.

Zhongwei Shen is with School of Electronic and Information Engineering, Suzhou University of Science and Technology, 215009, Suzhou, China

1 Hui Li and Yongbiao Xiao contributed equally to this work and should be considered co-first authors. *(* Corresponding author: Hui Li. Email: lihui.cv@jiangnan.edu.cn.)*

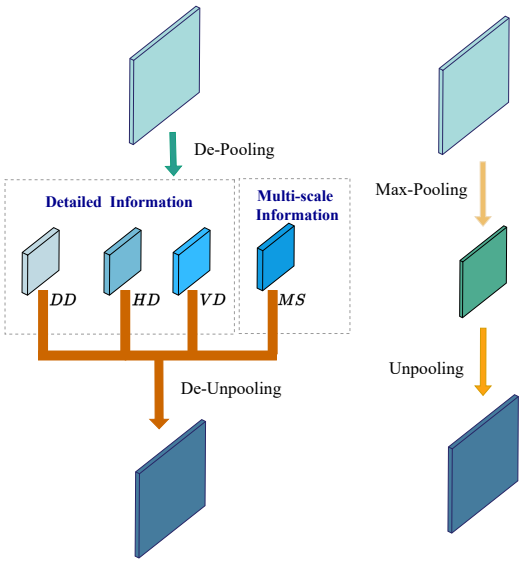


Fig. 1. Comparison between de-pooling based architecture (a) and max-pooling based model (b). Given the features, our de-pooling can preserve multi-scale features and detail features simultaneously, but the max-pooling operation can only extract multi-scale feature information

manually designed complex fusion strategies are only suitable for specific scenarios.

With the increase of multi-modal datasets, it is possible to design the end-to-end fusion network and the learnable fusion strategies. The generative adversarial network (GAN) [29]based methods are ideal for unsupervised fusion task. FusionGAN [24] is the first method to apply GAN into image fusion tasks. In FusionGAN, the generator aims to generate fused images with salient features and visible gradients, and the discriminator aims to force the fused image to preserve more texture details from visible image. However, the training of GAN networks is unstable and this method does not focus on downstream tasks. Therefore, in order to fit the high-level vision tasks, such as semantic segmentation and object detection, SeAFusion [30] cascades the image fusion module and semantic segmentation module, and simultaneously designs a gradient residual dense block to enhance fine-grained spatial details. Furthermore, more and more networks focus on the fusion of nighttime images. DIVFusion [31] designs two modules to remove the illumination degradation and enhance the contrast and texture details of the fused features respectively. It reasonably lights up the darkness and facilitates complementary information aggregation. Moreover, most methods only focus on the features from the source images and ignore the intermediate output of the network itself. MUFusion [32] introduces a novel memory unit architecture that mainly utilizes intermediate fusion results obtained during training to further supervise the fused images. SGFusion [33] as an end-to-end network can be applied to various fusion tasks. The network uses dual-guided encoding, image reconstruction decoding and saliency detection decoding to simultaneously extract feature maps and saliency maps at different scales from the image. In addition, with the development of hyperspectral (HSI) and multispectral (MSI) image fusion, Dian et al. [34] propose a zero-shot learning (ZSL) method for HSI sharpening and achieve excellent experimental results.

Although current end-to-end fusion methods can produce considerable results in most scenarios, there are still some drawbacks. Firstly, the lack of scale transformation in most fusion networks will lead to the inability to change the spatial scale of deep features. Secondly, some methods that introduce pooling to enrich features by scale transformation will bring a lot of information loss and artifacts. Finally, this kind of networks are not easy to train and the results are relatively unstable.

To extract the multi-scale deep features, the existing fusion networks only utilize the pooling operation, such as max-pooling (Figure 1 (b)). Although it can extract multi-scale feature information, a lot of features are omitted which are important for fused image reconstruction. To address the above issue, we propose a novel decomposition pooling (de-pooling, Figure 1 (a)). Compared with max-pooling, de-pooling can extract multi-scale information without information omits. In addition, as shown in visualization results in Figure 2, although most methods are valid by using max-pooling operations, they are not able to solve the information loss and artifacts. On the contrary, our method (Figure 2 (d)) avoids these drawbacks while enhancing the texture details.

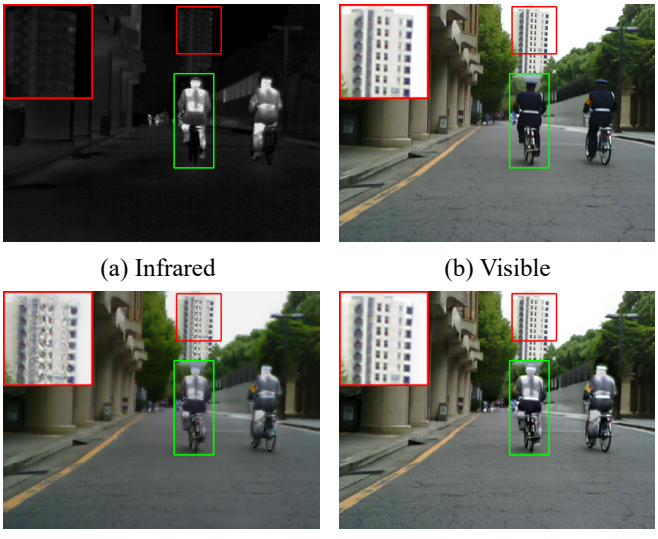


Fig. 2. Comparison between visualization results of the max-pooling based fusion method and the de-pooling (ours) based fusion method.

The previous wavelet based convolutional networks [35], [36], [37], [38] do not change the form of the kernel and they are mostly based on the kernel of the traditional wavelet transform. Compared with it, our method improves each kernel (Multi Scale, Vertical Direction, Horizontal Direction, Diagonal Direction) based on the Prewitt operator to capture multi-scale information (F_{MS}) and vertical, horizontal and diagonal (F_{VD}, F_{HD}, F_{DD}) texture details (details in III-A). In addition, the receptive field of these methods is mostly 2×2, but the receptive field of the 4×4 we use is larger, which can maintain fine-grained features (details in IV-F1).

In our fusion framework, we propose an innovative de-pooling operation that is applied within the encoder. This operation serves to extract multi-scale features and detail information at each scale, thereby enhancing the representational capabilities of the model. Regarding our fusion strategy, we incorporate a spatial attention model [22] to seamlessly combine complementary information. This approach is not only conceptually elegant but also highly efficient in practice. For the decoder, it operates with a wealth of multi-scale information and intricate texture details derived from the source images, enabling the faithful reconstruction of features. Furthermore, our decoder employs a novel technique known as inverse de-pooling (de-unpooling) during the up-sampling of deep features. This technique facilitates the integration of the fused detail information, contributing to the richness and accuracy of the reconstructed features. The utilization of inverse de-pooling underscores the unique and effective nature of our method in preserving and enhancing information throughout the fusion process. Due to its property of minimal information loss, the de-unpooling is able to completely reconstruct images without any post-processing step, which is crucial for image generation. The major contributions of this study are summarized as follows:

- A novel pooling operation named de-pooling is designed. It can extract multi-scale features and simultaneously preserve the detail information at each scale, which is helpful for the image fusion task.
- A simple and efficient fusion network is proposed based on this framework, and results can be achieved even only with a simple fusion strategy.
- Our fusion method can enhance the performance of high-level vision tasks (*e.g.*, object detection), achieve robustness to noise, and improve operational efficiency without requiring additional computational resources, while showcasing its distinctive superiority through qualitative and quantitative experiments on multiple image fusion benchmarks.

The rest of this paper is organized as follows: Section II discusses related work on image fusion. In Section III, we present the details of the proposed fusion method. Section IV shows the experimental results and compares them with some state-of-the-art methods. Finally, we present the conclusions in Section V.

II. Related Works

A. Multi-scale transform based fusion methods

Multi-scale decomposition is the most common feature extraction technique, such as Laplacian pyramid, wavelet transform [39], complex wavelet transform [13], curvelet [12], contourlet [40], shearlet [41], etc. Its basic step is first to obtain a multi-scale representation of the input image using a multi-scale transform. Then, the multi-scale representations of different images are fused according to specific fusion rules to obtain the fused multi-scale representation. Finally, the fused image is obtained by multi-scale inverse transform. In addition, there are some specific fusion rules, including choose-max [42], weighted-average [43], optimization based method [44], element-wise addition [45] and so on.

In general, most non-deep learning methods are time-consuming. The manually designed feature strategies cannot handle complex situations.

B. Fusion methods based on deep learning

Due to the powerful feature representation ability, deep learning is widely used in most computer vision tasks. A typical auto-encoder based fusion method is DenseFuse [21]. It uses dense blocks during encoding to extract features, where the output of each layer is adapted as the input to the next layer. Finally, the fusion image was reconstructed by the fusion strategy and the decoding network. In addition, AEFusion [28] captures long-range semantic information while extracting multi-scale features, combining with Axial-attention. A Transformer based auto-encoder presented, it uses self-supervised multi-task learning, called TransMEF [46], which enables the network to deal with both local and global information at the same time. SEDRFuse [47] uses a symmetric encoder-decoder and AUIF [48] converts the two models into a basic encoder and a detail encoder. But these cannot avoid manually designing complex fusion strategies.

Therefore, other researchers focus on exploring the end-to-end CNN based image fusion networks, which rely on the complex network structure and the carefully designed loss function. A representative general fusion network is IFCNN [49]. In addition, PSTL [50] deploys content branch and detail branch to extract characteristics and PIAFusion [51] based on illumination-aware is proposed to take the illumination factor into account in the modeling process. APWNet [52] proposes a simple yet efficient network that adaptively learns pixel-by-pixel weights for image fusion, and the joint optimization is effective for improving fusion quality and detection accuracy. However, their enhancement of details only through convolutional networks is not as good as our decomposition pooling.

In addition, GAN-based fusion methods are constantly innovating. AT-GAN [53] combines intensity attention modules and semantic transformation modules, which can effectively extract key information from multi-modal images. However, GAN obviously has some limitations in dealing with multi-scale information and details. TGFuse [54] innovatively combines Transformer and GAN. Although they can effectively learn the global fusion relationship, the extraction of details is not as good as our decomposition pooling.

In recent years, many researchers have proposed Transformer-based infrared and visible image fusion algorithms. SwinFusion [55] proposes a general image fusion framework based on cross-domain long-range learning and Swin Transformer. SwinFuse [56] constructs a full attention feature encoding backbone to model long-range dependencies. This is a pure Transformer network. DATFuse [57] uses a dual attention transformer which can extract important features and preserve global complementary information. Although the training of Transformer models requires high computer performance, more advanced Transformer-based

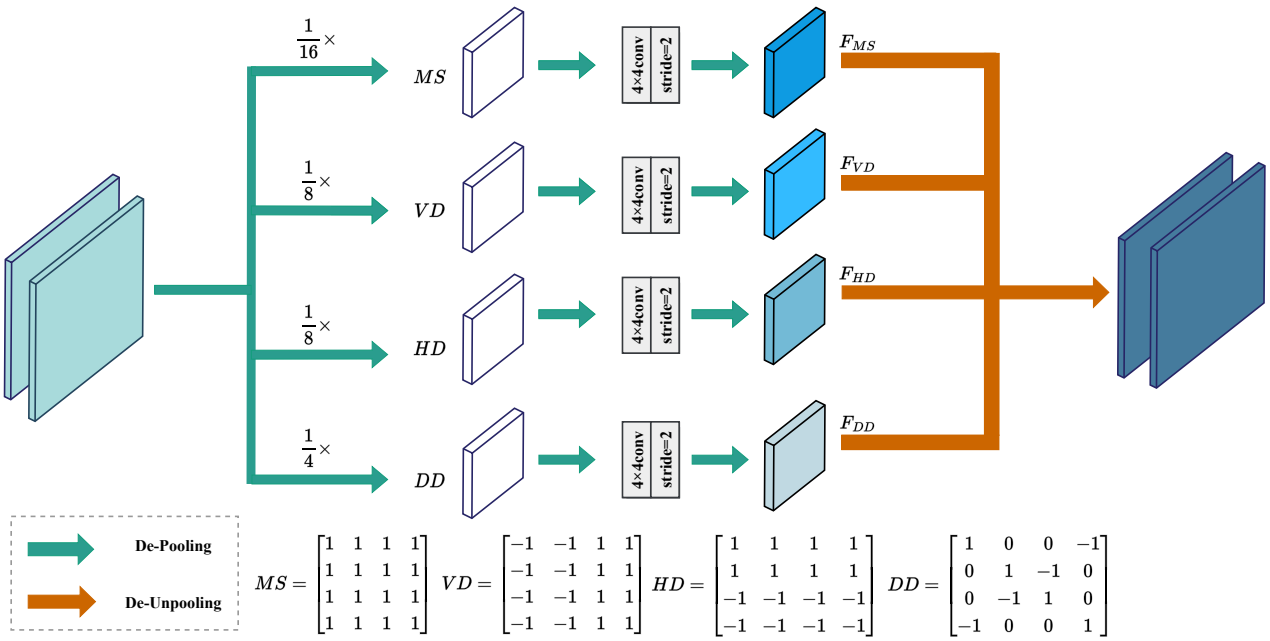


Fig. 3. The proposed module using decomposition pooling and unpooling. We employ 4×4 convolution kernel with stride 2 to extract multi-scale features F_{MS} and detail features (F_{VD}, F_{HD}, F_{DD}). Only the multi-scale component MS passes to the encoding layer and the detail components $\{VD, HD, DD\}$ are skipped to the corresponding decoding layer after feature fusion. In the decoder, the components are aggregated by the de-unpooling.

algorithms may be proposed in the future due to the great scientific research potential of Transformers.

On the whole, most fusion algorithms based on deep learning do not change the spatial scale, and the feature richness is relatively insufficient. On the other hand, the method with multi-scale features only employs general pooling operations, which leads to the loss of information when extracting features. However, the lost information is beneficial to image generation and reconstruction, especially in the fusion task. This requires us to find other methods to combine with deep learning based methods.

C. Combining multi-scale transform and deep learning

When people pay too much attention to deep learning methods, they will ignore the ideas of non-deep learning methods. Therefore, more and more researchers have proposed a combination of non-deep learning and deep learning methods, which can use the excellent ideas of non-deep learning methods and the powerful framework of deep learning.

Chen [58] et al. propose a convolutional encoder–decoder image fusion network based on wavelet transform. They use max-pooling and up-sampling in the sub-network. The method proposed by Wang [59] et al. uses convolutional neural networks in the discrete wavelet transform domain. They introduce the low-frequency information and high-frequency information of DWT into CNN-high and CNN-low networks to extract features, and similarly, they also use max-pooling operation in the network. Xu [60] et al. proposed a multi-scale feature pyramid network based on activity level weight selection for infrared and visible image fusion. In this method, they use max-pooling and up-sampling convolution blocks to extract features and restore image details. Chao [61] et al propose the discrete stationary wavelet transform (DSWT) to extract the high-frequency and low-frequency components of the image, and then they put them into radial basis function neural network (RBFNN) to extract features. Finally, the methods perform the inverse operation to reconstruct the image.

Although these methods can extract effective information in some scenarios, most of them use max-pooling to extract features, which leads to information loss and introduces noise.

The wavelet-based pooling work [37] is proposed as an alternative to traditional neighborhood pooling. However, they reduce the feature dimension by discarding the first-level components, while our method needs to utilize all the components. In LDWPooling [38], although the proposed Learning Discrete Wavelet Pooling can be very good for image recognition and detection, it is not good enough for fine details processing, and the post-processing stage requires fine selection of key features and representative features, which is too complex.

In contrast, most pooling operations either cause information loss or extract part of the key features, which cannot adapt to complex multi-modal image fusion tasks. Our proposed de-pooling fully extracts multi-scale information while preserving texture details at each scale. In addition, de-unpooling can completely reconstruct the features without any post-processing operations and finally generate a fused image containing rich information.

III. Proposed Fusion Method

In this section, the proposed fusion method will be presented. Firstly, we give the details of decomposition pooling. Then, the model architecture is introduced. Finally, the details of training phase will be given.

A. Decomposition Pooling

Compared with the common pooling operation, the decomposition pooling contains four kernels, $\{MS, VD, HD, DD\}$,

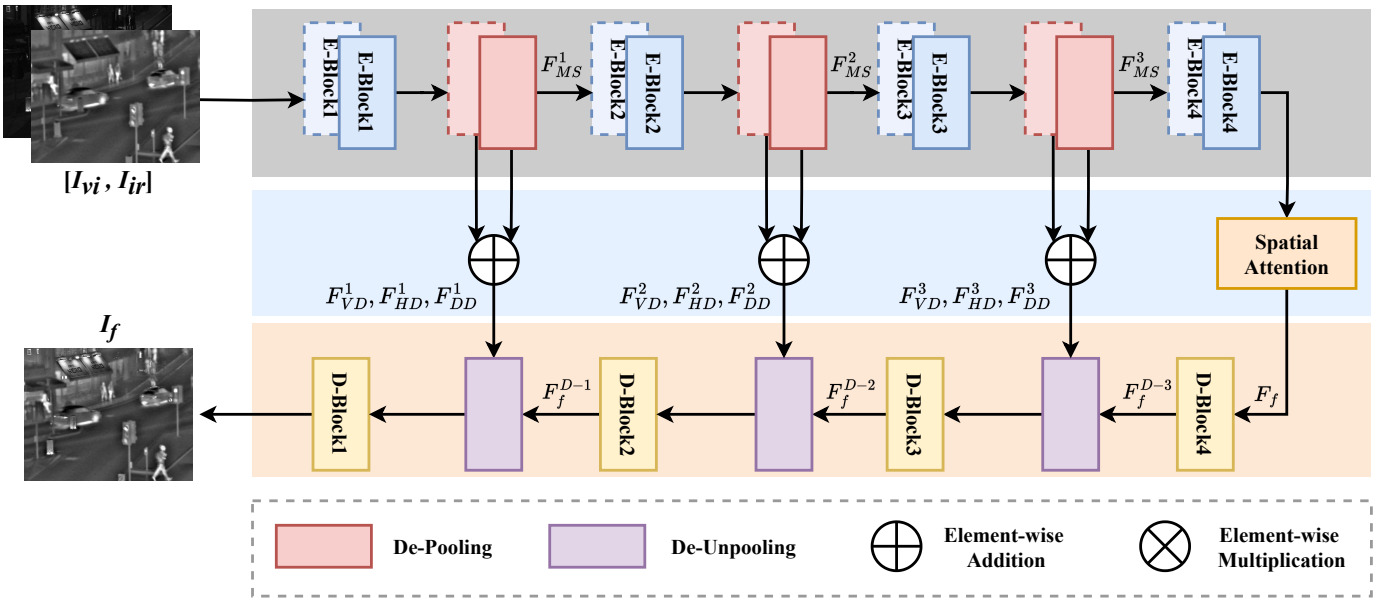


Fig. 4. The architecture of the proposed method. A pair of encoder and decoder are at same scale. The encoder includes four convolutional blocks and three de-pooling, and the decoder consists of four convolutional blocks and three de-unpooling operations. De-unpooling is the inverse operation of de-pooling. The fusion layer uses spatial attention model (details in III-B2) and the detail information $\{F^i_{VD}, F^i_{HD}, F^i_{DD}\}$ extracted by ith de-pooling employs addition strategy. F^i_{MS} represents the multi-scale features extracted after the ith de-pooling. F^{D-i}_f denotes the feature information reconstructed by ith de-unpooling together with the detail information.

which represent four feature spaces respectively. As shown in Figure 3, the proposed de-pooling also contains four parts, the multi-scale component MS captures global multi-scale information, while the detail components VD, HD and DD extract vertical, horizontal, and diagonal texture details.

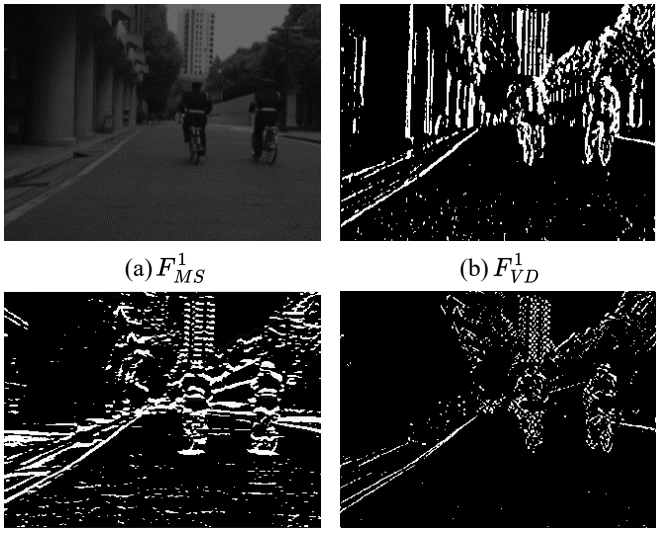
Inspired by the Prewitt operator, our four feature spaces are formulated as following:

$$MS = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad VD = \begin{bmatrix} -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \tag{1}$$

$$HD = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 \end{bmatrix} DD = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

In our proposed fusion network, we replace the common max-pooling layer with our de-pooling. The multi-scale components are passed to the encoding layer through convolutional blocks and decomposition pooling while de-unpooling utilizes the detail components of the source images for feature reconstruction. The four feature maps after the first de-pooling are shown in Figure 5. Specifically, the MS component extracts multi-scale information, and the detail components VD, HD and DD extract texture details in three directions (vertical, horizontal and diagonal).

Our de-unpooling has the property of minimal information loss and can reconstruct the images exactly. On the other hand, since max-pooling does not have its exact inverse, the networks proposed by Chen [58] et al. and Wang [59] et al. cause greater information loss during the reconstruction process.



(a) F^1_{MS} (b) F^1_{VD} (c) F^1_{HD} (d) F^1_{DD}

Fig. 5. Visualization results of four feature maps after the first de-pooling. The global multi-scale information is obtained by multi-scale component MS, and the detail components VD, HD, DD capture vertical, horizontal, and diagonal texture details.

B. Network Architecture

In order to fully extract the salient information and texture details from the multi-modal image and reconstruct the structure information of the source image, a novel fusion network architecture based on de-pooling is proposed. Our network architecture is composed of three parts: encoder, fusion layer, and decoder. The architecture is shown in Figure 4. The input infrared and visible images are denoted as I_{vi} and I_{vi}. Note that the input images are pre-registered. The source images are processed by de-pooling to extract multi-scale information and

simultaneously preserve detail information. In addition, the spatial attention model serves as our fusion strategy and each scale of the detail components corresponds to the addition. Finally, the extracted features are used to reconstruct the image through de-unpooling.

1) Encoder: The proposed encoder is shown in Figure 4 (grey box). The solid and dashed lines in the Block represent the same structure and the same parameters. It has four convolutional blocks and three de-pooling operations. The proposed de-pooling can generate two pieces of information simultaneously: multi-scale structure information and detail information (contains vertical, horizontal, and diagonal texture details). Within the encoder, we employ de-pooling to extract multi-scale information while preserving texture details for later reconstruction, which can generate better images.

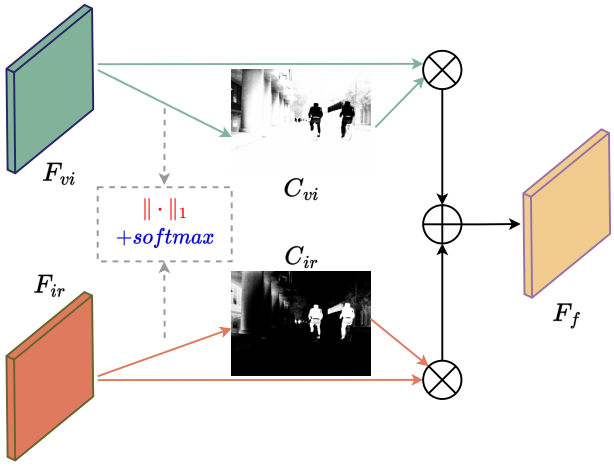


Fig. 6. Procedure of spatial attention fusion strategy.

2) Fusion Strategy: **Multi-scale Features Fusion Strategy:** A well-designed fusion strategy is the key to the image fusion task. In this work, we introduce spatial attention model [22], [18] into our method to fuse multi-scale deep features, in which the attention weights are calculated adaptively. The procedure for obtaining the spatial attention model is shown in Figure 6.

Given the multi-scale deep features F_{ir} and F_{vi} which are extracted by the last block of encoder ("E-Block4") from infrared and visible images respectively, the weighting maps which are calculated by l_1-norm and soft-max operator can be calculated by:

$$C_\omega(x,y) = \frac{\|F_\omega(x,y)\|_1}{\sum\limits_{i\in\{ir,vi\}} \|F_i(x,y)\|_1}, \quad \omega \in \{ir, vi\} \tag{3}$$

where $\|\cdot\|_1$ denotes l_1-norm. The l_1-norm of F_ω can be the activity level measure of feature maps. The corresponding weight maps can be obtained by the above division operation, as shown in Figure 6 (C_{vi} and C_{ir}). (x, y) indicates the corresponding position in multi-scale features (F_{ir} and F_{vi}) and weighting maps (C_{ir} and C_{vi}).

Then, $\beta_{ir}(x, y)$ and $\beta_{vi}(x, y)$ denote the weighted deep features, which are calculated by the following:

$$\beta_\omega(x,y) = C_\omega(x,y) \times F_\omega(x,y), \quad \omega \in \{ir, vi\} \tag{4}$$

Finally, the fused feature map F_f is calculated by adding enhanced deep features, the formulation is shown in the following equation:

$$F_f = \sum_{\omega\in\{ir,vi\}} \beta_\omega(x,y) \tag{5}$$

Detail Features Fusion Strategy: In addition, each scale corresponds to the detail information. Since the detail components are sparse, which are also very important for the generated image texture. Thus, in our framework, we are willing to preserve it as much as possible. The detail features fusion strategy is formulated as follows:

$$F_m^i = F_m^{ir} + F_m^{vi}, \quad m \in \{VD, HD, DD\} \tag{6}$$

Where F_m^{ir} and F_m^{vi} represent the texture details of infrared and visible images in vertical, horizontal, and diagonal directions, respectively. F_m^i denote the final generated detail information through ith pooling layer in the three directions.

3) Decoder: As shown in Figure 4 (yellow box), the decoder reflects the encoder structure with four convolutional blocks and three de-unpooling. For better reconstruction features, we adopt summation for unpooling. Specifically, our de-unpooling performs the inverse operation of the processed detail features and multi-scale features output by de-pooling from the corresponding scales. Therefore, the decoder with de-unpooling is able to completely reconstruct the image features.

TABLE I
NETWORK ARCHITECTURE OF OUR METHOD. **INPUT** AND **OUTPUT** DENOTE THE NUMBER OF CHANNELS IN THE CORRESPONDING FEATURE MAPS.

	Blocks	Layers	Kernel	Input	Output	Activation
Encoder	E-Block1	Layer1	1×1	1	1	ReLU
		Layer2	3×3	1	64	ReLU
		Layer3	3×3	64	64	ReLU
	E-Block2	Layer1	3×3	64	128	ReLU
		Layer2	3×3	128	128	ReLU
	E-Block3	Layer1	3×3	128	256	ReLU
		Layer2	3×3	256	256	ReLU
		Layer3	3×3	256	256	ReLU
		Layer4	3×3	256	256	ReLU
	E-Block4	Layer1	3×3	256	512	-
Decoder	D-Block4	Layer1	3×3	512	256	-
	D-Block3	Layer1	3×3	256	256	ReLU
		Layer2	3×3	256	256	ReLU
		Layer3	3×3	256	256	ReLU
		Layer4	3×3	256	128	ReLU
	D-Block2	Layer1	3×3	128	128	ReLU
		Layer2	3×3	128	64	ReLU
	D-Block1	Layer1	3×3	64	64	ReLU
		Layer2	3×3	64	1	ReLU

C. Training Phase

Our training phase is based on the auto-encoder training strategy, and the fusion strategy is discarded in this phase. With this strategy, the encoder is able to extract multi-scale structural information and texture details from input images, the decoder can well reconstruct the input from the extracted information. The training framework is shown in Figure 7, and the network architecture is shown in Table I.

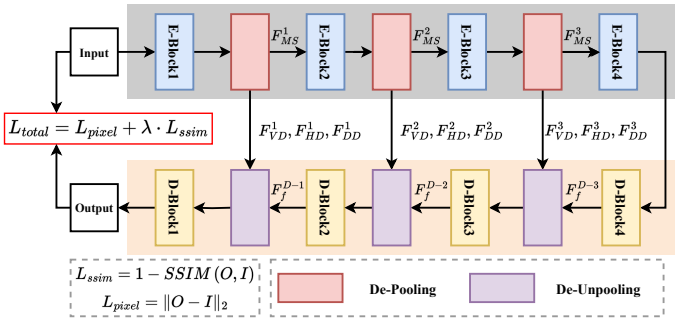


Fig. 7. The framework of training process. In this process, we train an auto-encoder network without fusion strategy.

In order to reconstruct the input image more precisely, the loss function L_{total} is defined as follows:

$$L_{total} = L_{pixel} + \lambda L_{ssim} \tag{7}$$

where L_{pixel} and L_{ssim} indicate the pixel loss and structure similarity (SSIM) loss, respectively. λ denotes the tradeoff value between L_{pixel} and L_{ssim}.

L_{pixel} is calculated by the following equation:

$$L_{pixel} = \|O - I\|_F^2 \tag{8}$$

where O and I indicate the output and input images, respectively. $\|\cdot\|_F^2$ is the l_2-norm. This loss function will make sure that the reconstructed image is more similar to the input image at the pixel level.

The SSIM loss L_{ssim} is formulated as follows:

$$L_{ssim} = 1 - SSIM(O, I) \tag{9}$$

where $SSIM(\cdot)$ denotes the structural similarity between the generated images and source images [62].

IV. Experimental Results and Analysis

A. Experimental Settings

In the training phase, we randomly choose 40000 images from MS-COCO [63] dataset to train our auto-encoder network, all of them are resized to 256×256. The learning rate is set as 1×10^{-4}. The batch size and epochs are set to 4 and 4, respectively. The hyperparameter λ in Equation 7 is set as 100. All the involved experiments are conducted on an NVIDIA RTX 3090Ti GPU and Intel Core i7-10700 CPU. To fuse the RGB images, firstly, the visible images are converted to the YCbCr color space. Then, the Y channel of the visible images and the infrared images are fused by the proposed method. Finally, the fused image is converted back to the RGB color space by concatenating the Cb and Cr channels of the visible images.

To comprehensively evaluate the proposed algorithm, we perform qualitative and quantitative experiments on the MSRS dataset [51] with all the 361 image pairs, the LLVIP dataset [64] with randomly selected 389 image pairs and the TNO dataset [65] with randomly selected 16 image pairs.

To evaluate the fusion performance, we choose nine typical and state-of-the-art fusion methods, including DenseFuse [21], FusionGAN [24], SwinFusion [55], SwinFuse [56], U2Fusion [66], AUIF [48], CUFD [67], MUFusion [32] and AEFusion [28]. The implementations of these approaches are publicly available.

Five quality metrics are utilized for quantitative comparison between our fusion method and other existing fusion methods, including standard deviation (SD), visual information fidelity (VIF), average gradient (AG), the sum of correlations of differences (SCD) and entropy (EN). SD reflects the visual effect of the fused image. VIF measures the information fidelity of the fused image. AG quantifies the gradient information of the fused image and represents its detail and texture. SCD reflects the level of correlation between the information transmitted to the fused image and corresponding source images. EN is used to represent the image detail retention. Moreover, a fusion algorithm with larger SD, VIF, AG, SCD and EN indicates better fusion performance.

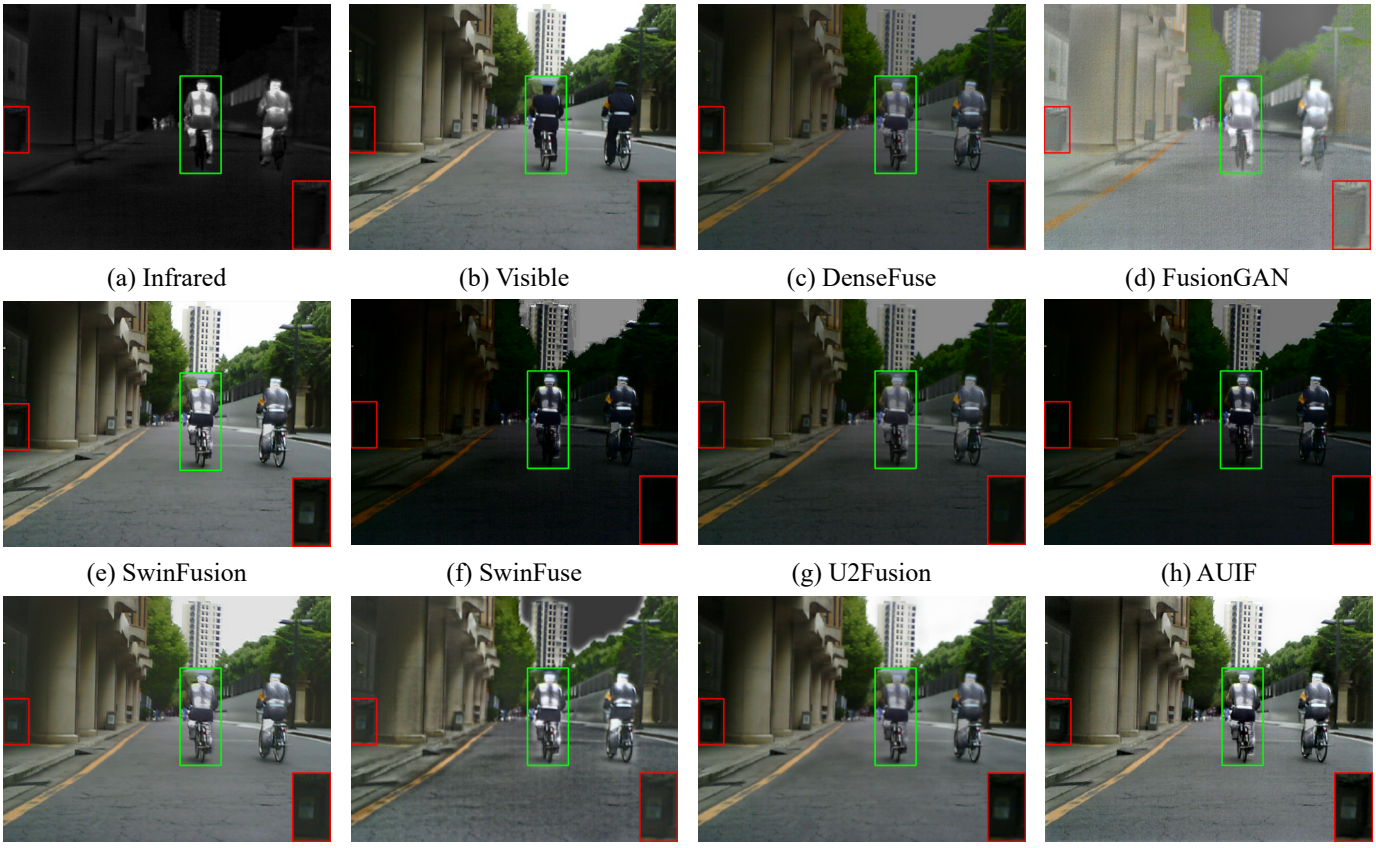
B. Comparative experiments

In order to fully demonstrate the distinctive superiority of our approach, we compare our method with other nine SOTA methods on MSRS, LLVIP and TNO datasets. [1]

1) Fusion Results on MSRS Dataset: Qualitative experiments performed on the MSRS dataset are shown in Figure 8. As shown in the red highlighted regions, SwinFuse, U2Fusion and AUIF can barely preserve the detail information of the visible image. In addition, FusionGAN suffers from severe spectral contamination, which degrades the visual effect of the fused image. Although DenseFuse preserves part of the texture details, it weakens salient features, resulting in an overall darker effect. In the green box, we can find that MUFusion blurs the edges of targets and introduces a lot of artifacts and noise into results in some regions, causing visual conflicts. In contrast to our results, although CUFD and AEFusion integrate the texture information of visible images with salient target information in infrared images to some extent, they produce abundant artifacts that blur edges and details, making the overall effect is inferior to ours. Moreover, SwinFusion overextracts the infrared features of characters, resulting in an overexposed scene. In general, only our method successfully maintains the structure intensity and preserves the textures, thanking to our pooling and unpooling operation.

The quantitative results of seven metrics on the MSRS dataset are presented in Table II. The subjective comparison in Figure 8 presents the superiority of our method in SD and VIF metrics (following SwinFusion by a narrow margin), indicating that our results have high contrast, maximization of

[1]For more experiments, please refer to our supplementary material.



(a) Infrared (b) Visible (c) DenseFuse (d) FusionGAN

(e) SwinFusion (f) SwinFuse (g) U2Fusion (h) AUIF

(i) CUFD (j) MUFusion (k) AEFusion (l) Ours

Fig. 8. Qualitative comparison of our method with nine state-of-the-art methods on the MSRS dataset.

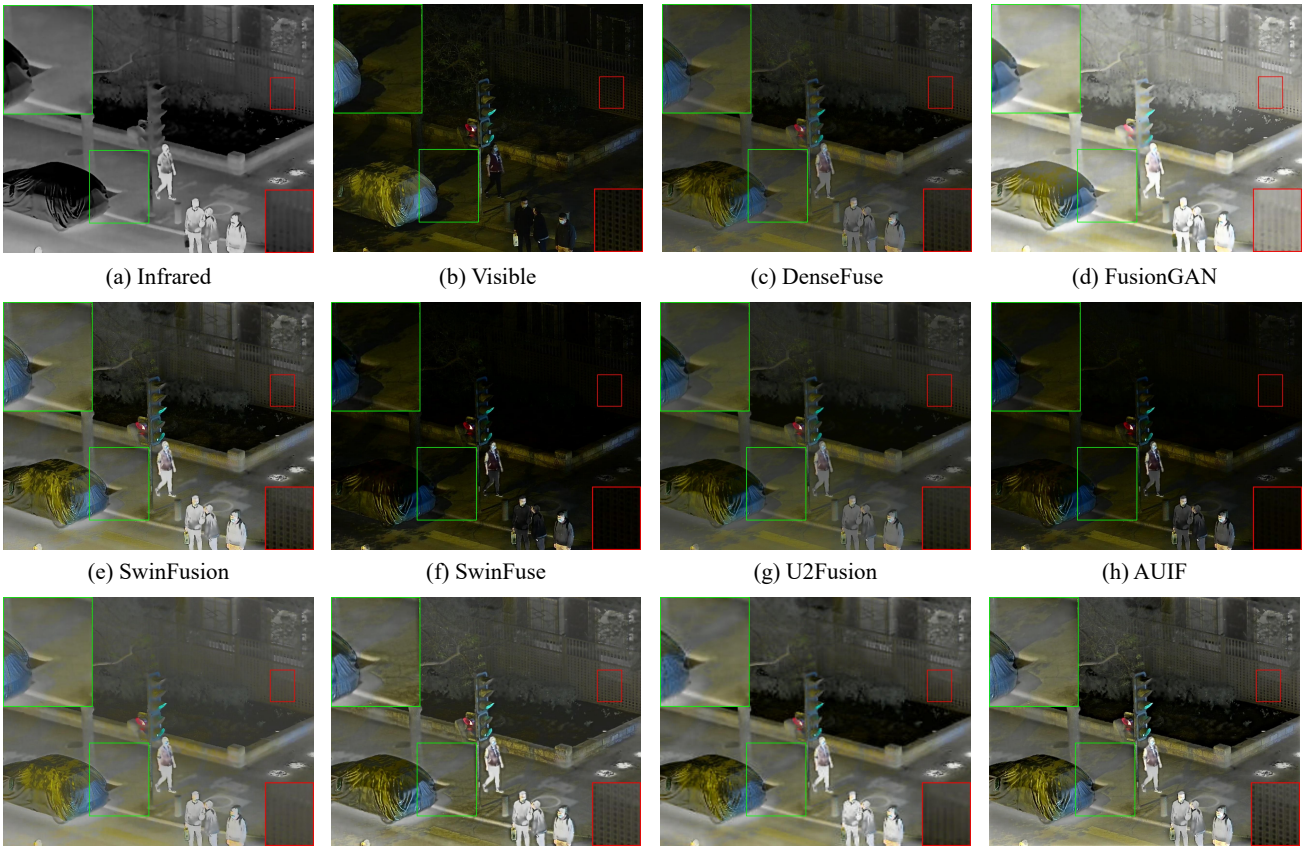
TABLE II
QUANTITATIVE RESULTS ON 361 IMAGE PAIRS FROM THE MSRS DATASET. (**BOLD**: BEST, RED: SECOND BEST, BLUE: THIRD BEST)

Methods	SD	VIF	AG	SCD	EN
DenseFuse	7.4237	0.6999	2.0873	1.2489	5.9340
FusionGAN	7.1758	0.8692	3.1193	0.3129	5.9937
SwinFusion	8.3928	**1.0109**	3.5434	1.6908	6.6196
SwinFuse	4.9246	0.4102	1.9673	1.0129	4.4521
U2Fusion	6.8217	0.5863	2.0694	1.2955	5.5515
AUIF	5.2622	0.3981	1.8238	1.0639	4.6460
CUFD	7.6384	0.6488	2.9003	1.2379	6.0652
MUFusion	6.9233	0.6086	3.1474	1.2548	5.9682
AEFusion	8.2104	0.8548	2.6968	1.4564	6.5374
Ours	**8.4779**	0.9508	**4.0498**	**1.7159**	**6.7697**

the information fidelity and satisfying visual effects, which is consistent with the human visual system. For the AG metric, the first ranking of our method illustrates that our fusion results contain more edge information. From the results, we can see that our framework ranks first in SCD and EN metrics, which demonstrates that our fusion results retain realistic and valid information correlated with the source images.

2) Fusion Results on LLVIP Dataset: The visualization results of different methods on the LLVIP dataset are shown in Figure 9. The selected nighttime image can demonstrate the superiority of our approach. As can be seen from the red box, SwinFuse and AUIF fail to retain detail information of the fence. Although FusionGAN, CUFD and AEFusion maintain the infrared salient features, the background regions are all affected by varying degrees of spectral contamination, resulting in blurring detail information. In addition, DenseFuse and U2Fusion weaken the infrared features, making the overall images darker and affecting human visual perception. Only our method, SwinFusion and MUFusion effectively preserve the texture details of the fence. As observed from the green box, SwinFusion and MUFusion do not deal well with the details of branch shadows. They introduce artifacts and almost blend in with the ground without a sharp distinction. The same goes for other methods except our method. Our method does a better job in terms of significant target maintenance and details preservation than other methods.

The comparative subjective results of different methods on the LLVIP dataset are shown in Table III. Our method achieves the best results on 3 of these 5 metrics. The highest performance on AG proves that our fusion results are able to preserve rich gradient information. Meanwhile, the best results in SCD and EN metrics mean that our fusion results are highly consistent with the source images and can generate more realistic images. The SD metric of our method ranks second because FusionGAN introduces more noise and artifacts into the fused image. Moreover, we can also observe that FusionGAN in the VIF metric is optimal, but the texture information is lost in visualization, resulting in blurred details



(a) Infrared (b) Visible (c) DenseFuse (d) FusionGAN

(e) SwinFusion (f) SwinFuse (g) U2Fusion (h) AUIF

(i) CUFD (j) MUFusion (k) AEFusion (l) Ours

Fig. 9. Qualitative comparison of our method with nine state-of-the-art methods on the LLVIP dataset.

TABLE III
QUANTITATIVE RESULTS ON 389 IMAGE PAIRS FROM THE LLVIP DATASET. (**BOLD**: BEST, RED: SECOND BEST, BLUE: THIRD BEST)

Methods	SD	VIF	AG	SCD	EN
DenseFuse	9.2963	0.7503	2.6714	1.2109	6.8287
FusionGAN	**10.0823**	**1.0263**	2.1706	0.4276	7.1741
SwinFusion	9.6297	0.9807	4.3851	1.5678	7.3466
SwinFuse	7.5469	0.6290	2.9580	1.1840	6.0825
U2Fusion	9.4256	0.7212	2.3685	1.3114	6.7588
AUIF	7.5433	0.5877	2.8790	1.1667	6.1555
CUFD	9.1701	0.7187	2.5198	1.0360	6.8448
MUFusion	8.7452	0.7875	3.5412	1.0975	6.9242
AEFusion	9.8400	0.6302	2.0397	1.3526	7.2764
Ours	9.9393	0.9725	**4.7475**	**1.6155**	**7.4398**

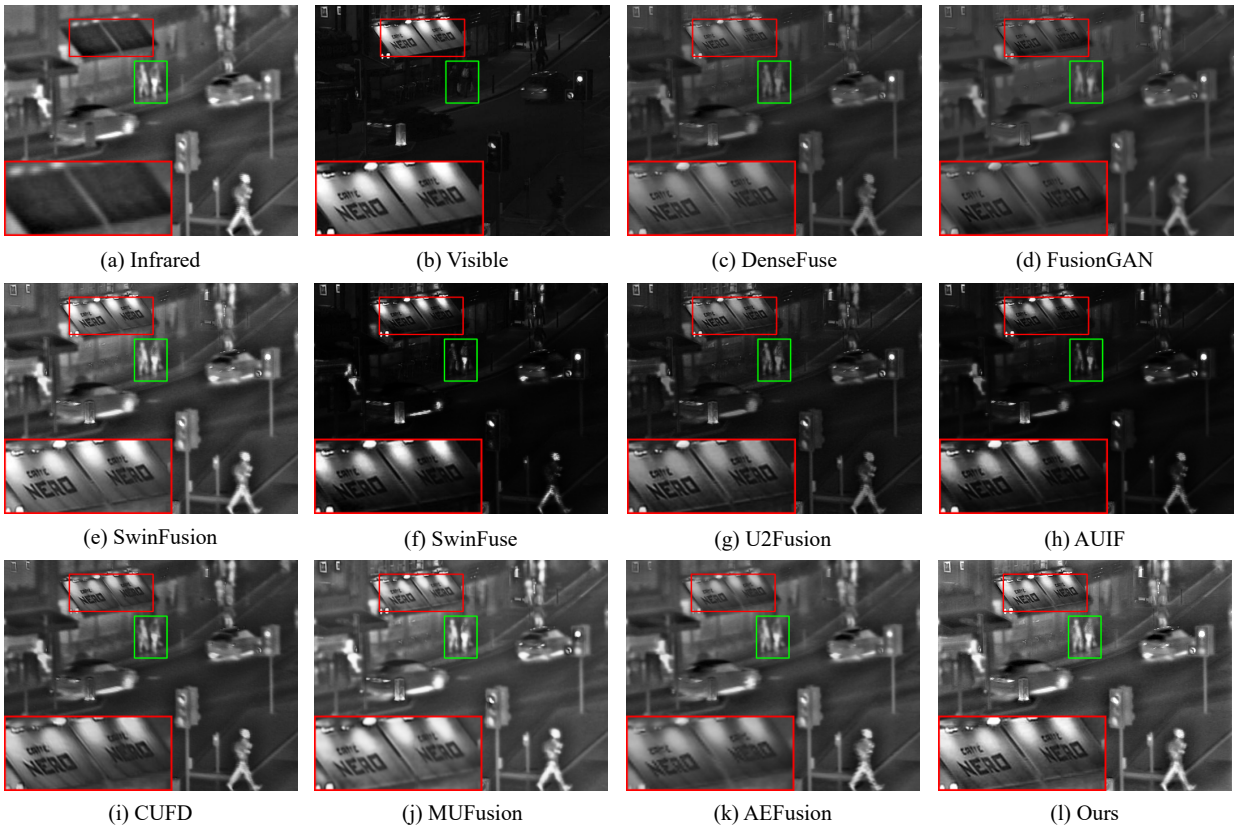
and image distortion.

3) Fusion Results on TNO Dataset: The qualitative comparisons of different algorithms on the TNO dataset are presented in Figure 10. As shown in the green boxes, DenseFuse, SwinFuse, U2Fusion and AUIF weaken the salient target. In addition, it can be seen from the red box that the overall images of SwinFuse and AUIF are darker than others and are unable to preserve the texture details of the visible images. For FusionGAN and AEFusion, they blur the sharp edges of visible images. In addition, CUFD and MUFusion introduce many artifacts into the fused images, resulting in worse visual perception. Only our method and SwinFusion are able to maintain structure intensities while preserving texture details.

TABLE IV
QUANTITATIVE RESULTS ON 16 IMAGE PAIRS FROM THE TNO DATASET. (**BOLD**: BEST, RED: SECOND BEST, BLUE: THIRD BEST)

Methods	SD	VIF	AG	SCD	EN
DenseFuse	9.2203	0.7349	3.8804	1.6300	6.8256
FusionGAN	8.1234	0.6197	2.8120	1.1911	6.4629
SwinFusion	9.5370	**0.8907**	5.3161	1.6855	7.0270
SwinFuse	9.2633	0.7982	5.5986	1.6882	6.9484
U2Fusion	9.3869	0.7200	5.4456	1.6406	6.9395
AUIF	9.2805	0.7482	5.2820	**1.7537**	7.0402
CUFD	9.4136	0.8781	4.5178	1.4135	7.0743
MUFusion	9.5379	0.7851	5.5756	1.5338	**7.3032**
AEFusion	9.4655	0.7803	3.4114	1.5744	7.0716
Ours	**9.6036**	0.7786	**6.0003**	1.5964	7.1583

We use quantitative metrics to measure the performance of different methods on the TNO dataset , which is shown in Table IV. The best results in AG and SD mean that our method captures abundant texture details and simultaneously achieves the best visual perceptual performance. Our method



(a) Infrared (b) Visible (c) DenseFuse (d) FusionGAN

(e) SwinFusion (f) SwinFuse (g) U2Fusion (h) AUIF

(i) CUFD (j) MUFusion (k) AEFusion (l) Ours

Fig. 10. Qualitative comparison of our method with nine state-of-the-art methods on the TNO dataset.

trails MUFusion by a narrow margin in the EN metric, but our method does not introduce artifacts and noise into the fusion results from visualizing results and also contains a lot of meaningful information transmitted from the source image to the fused image. However, for the VIF and SCD metrics, our method does not perform as well as on the other two datasets. This is justified since the source images in TNO dataset contain less detail information. Compared with other datasets (with RGB images), the visible images in TNO mainly have better infrared salient features, while texture details are not obvious. However, de-pooling prefers to preserve the texture details and the intensity of source images. Thus, in TNO dataset, our proposed method achieves comparable fusion performance.

In conclusion, both qualitative and quantitative results demonstrate that our method achieves better/comparable fusion performance. Moreover, our method has obvious advantages in preserving texture details and maintaining structure information, and can achieve a pleasing visual perception.

C. Comparison of the Efficiency

We compare the average running time taken by different methods in processing 361 pairs of infrared and visible images on the MSRS dataset. To make the test more convincing, the selected images are all of the same size (640×480). As shown in Table V, our method achieves the best performance on efficiency comparison with an average time of less than 0.03 s per image. Our approach significantly improves the efficiency without the need of the additional computational resources.

D. Detection Peformance

Although pixel-level image fusion is far less effective than feature-level and decision-level image fusion in improving the performance of high-level computer vision tasks such as detection, our method fully extracts multi-scale features and detail textures. Compared with single-modal images (infrared and visible images) and the latest fusion methods (MUFusion and AEFusion), our method can effectively improve the performance of high-level vision tasks such as object detection. To further demonstrate the superiority of the proposed method, we deploy the SOTA detection model YOLOv5 [68] to perform pedestrian detection on M3FD [69] dataset. We select the values of mAP@0.50-mAP@0.75 for comparison. Since all the accuracies after mAP@0.75 are below 0.2, the performance is low and the comparison is not worth much. As shown in Table VI, our method significantly outperforms the state of the art methods (MUFusion and AEFusion). In addition, our method also presents great superiority compared with single-modal images, which are infrared and visible images. This indicates that our method as pixel-level image fusion can also improve the performance of high-level computer vision tasks.

E. Noise Evaluation

It is difficult for existing instrumentation to avoid not introducing noise when acquiring images, so we conduct experiments on performance evaluation regarding to the noise to demonstrate the robustness of our method to noise. We add Gaussian noise with a variance of 0.5 to 361 pairs of infrared and visible images on MSRS dataset. As shown in

TABLE V
THE AVERAGE INFERENCE TIME (UNIT: SECOND) ON 361 PAIRS OF IMAGES FROM MSRS DATASET. (**BOLD**: BEST).

Method	DenseFuse	SwinFusion	FusionGAN	SwinFuse	U2Fusion	AUIF	CUFD	MUFusion	AEFusion	ours
Inference time	0.1813	0.0677	1.2371	0.2592	0.0342	0.1158	72.6157	0.7045	0.2244	**0.0230**

TABLE VI
PEDESTRIAN DETECTION PERFORMANCE FOR VISIBLE, INFRARED, AND FUSED IMAGES ON M3FD DATASET. (**BOLD**: BEST)

Methods	mAP@0.50	mAP@0.55	mAP@0.60	mAP@0.65	mAP@0.70	mAP@0.75
MUFusion	0.4658	0.4346	0.3994	0.3620	0.3132	0.2297
AEFusion	0.3984	0.3725	0.3437	0.3163	0.2754	0.2088
Inf	0.4997	0.4696	**0.4357**	**0.3949**	0.3371	0.2713
Vis	0.3877	0.3528	0.3217	0.2755	0.2131	0.1567
Ours	**0.5061**	**0.4733**	0.4311	0.3908	**0.3526**	**0.2722**

Table VII, our method can achieve the best 3 out of the 5 metrics, which means that our method performs better than the compared methods in dealing with noise. [2]

TABLE VII
QUANTITATIVE RESULTS WITH GAUSSIAN NOISE ON 361 IMAGE PAIRS FROM THE MSRS DATASET. (**BOLD**: BEST)

Methods	SD	VIF	AG	SCD	EN
DenseFuse	10.0092	0.2127	22.2801	0.3817	7.0382
FusionGAN	7.7310	**0.7564**	5.4857	0.2762	6.1292
SwinFusion	10.0769	0.1921	33.1547	0.3886	7.6917
SwinFuse	8.7608	0.2256	20.9663	0.4709	6.7314
U2Fusion	9.1773	0.2304	13.4323	0.5889	6.6359
AUIF	8.9527	0.2028	26.5962	0.3981	7.0671
CUFD	9.4456	0.1730	16.8999	0.3848	6.9860
MUFusion	9.1977	0.2487	14.7819	**0.6036**	6.9322
AEFusion	9.6373	0.1445	17.1373	0.4487	7.2088
Ours	**10.0782**	0.2281	**33.2012**	0.4229	**7.7133**

F. Ablation Studies

1) Analysis of the difference kernel of de-pooling: In our fusion framework, the proposed de-pooling uses 4×4 convolution kernels instead of 2×2 convolution kernels [70]. The fundamental difference in the size of the convolution kernel is the different receptive fields. The receptive field reflects the perception range of the current input feature map of the convolution kernel. If the range is small, the received information is one-sided and local. As the receptive field increases, more global information can be obtained, which is more conducive to the judgment of the current situation. Therefore, a large convolution kernel leads to a larger receptive field. However, blindly increasing the convolution kernel will bring the problem of increasing the amount of calculation. Therefore, it is necessary to choose the appropriate convolution kernel in deep learning.

As shown in Table VIII, the metrics of our 4×4 convolution kernel are better than the 2×2 convolution kernel in three metrics on the MSRS dataset and TNO dataset, and in four metrics on the LLVIP dataset. This fully demonstrates that the de-pooling we designed can better contain rich information and texture details, while achieving the best visual quality.

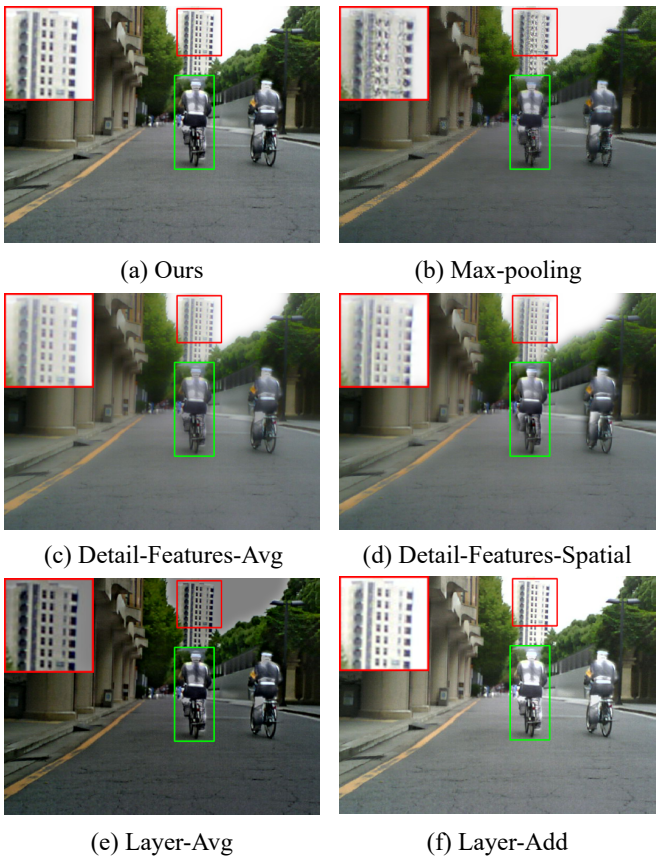
TABLE VIII
QUANTITATIVE COMPARISON OF 2×2 CONVOLUTION KERNELS AND 4×4 CONVOLUTION KERNELS IN MSRS, LLVIP AND TNO DATASETS.(**BOLD**: BEST)

Metrics	MSRS Dataset		LLVIP Dataset		TNO Dataset	
	2×2	4×4	2×2	4×4	2×2	4×4
SD	8.4648	**8.4779**	**9.9500**	9.9393	9.5748	**9.6036**
VIF	**0.9570**	0.9508	0.9682	**0.9725**	**0.7887**	0.7786
AG	3.9988	**4.0498**	4.7143	**4.7475**	5.9975	**6.0003**
SCD	**1.7358**	1.7159	1.6134	**1.6155**	**1.6394**	1.5964
EN	6.7486	**6.7697**	7.4315	**7.4398**	7.1394	**7.1583**

2) Max-pooling vs De-pooling: To demonstrate that our model does benefit from decomposition pooling, we compare fusion results using max-pooling. As shown in Figure 11 (b), although max-pooling can extract the salient features, it leads to degradation in fine details such as buildings (red box). In addition, the max-pooling introduces a lot of artifacts and noise, which is also because the max-pooling does not have its exact inverse. The proposed de-pooling can not only successfully preserves fine details and maintain structure information, but also has the characteristics of minimum information loss, allowing the network to completely reconstruct the signal. The quantitative results in Table IX show that our method outperforms max-pooling in all metrics

3) Analysis of detail features fusion strategy: In our method, the detail information (VD, HD, DD) extracted by de-pooling is added correspondingly, and then the de-

[2]For the visualization of results, please refer to our supplementary material.



(a) Ours (b) Max-pooling

(c) Detail-Features-Avg (d) Detail-Features-Spatial

(e) Layer-Avg (f) Layer-Add

Fig. 11. Visualized results of ablation studies. From (a) to (f): fused results of our method, fused results of using max-pooling, detail features with different directions using average strategy, detail features with different directions using spatial attention fusion strategy, fusion layer using average strategy and fusion layer using addition strategy.

unpooling is guided by this information to reconstruct features. To prove that the addition operation can better preserve edge information, we compare this strategy to the average strategy and spatial attention fusion strategy in this experiment. As shown in Figure 11 (c) and (d), average strategy and spatial attention fusion strategy for detail features will weaken the structure information and blur the edge of background regions in the red box. Moreover, these two kinds of strategies introduce a large number of artifacts and achieve a bad visual effect. It can be seen that the detail features with addition strategy can maintain structure intensities and strengthen edge information. The quantitative results are shown in Table IX and our method achieves the best results.

4) Analysis of multi-scale features fusion strategy: For the last layer of the encoder, we use the simple spatial attention fusion strategy to fuse multi-scale features. The spatial attention fusion strategy is utilized to fuse multi-scale deep features. In this ablation study, the average strategy and the addition strategy are conducted. As shown in Figure 11 (e) and (f), the average strategy weakens the background brightness of the visible image and leads to an overall darker image. However, the addition strategy will excessively enhance the salient features of the infrared image, resulting in visual conflict. It can be observed that the spatial attention fusion strategy can better integrate multi-scale features and achieve satisfying visual senses.

TABLE IX
THE AVERAGE VALUES OF THE FIVE OBJECTIVE METRICS OBTAINED WITH DIFFERENT ABLATION STUDIES ON MSRS DATASET. (**BOLD**: BEST, RED: SECOND BEST, BLUE: THIRD BEST)

Strategies	SD	VIF	AG	SCD	EN
Ours	8.4779	**0.9508**	**4.0498**	1.7159	6.7697
Max-Pooling	7.8781	0.7087	2.6391	1.3769	6.3626
Detail-Features-Avg	8.3962	0.6736	2.1711	1.3599	6.5538
Detail-Features-Spatial	8.4381	0.5920	2.3306	1.2897	6.5724
Layer-Avg	7.6953	0.8894	4.0208	1.4381	6.3852
Layer-Add	**8.6109**	0.8464	3.8939	**1.8974**	**6.8248**

The average values of the five metrics are shown in Table IX. Our method has better performances than using average strategy. Compared with the addition strategy, only VIF and AG metrics of our method are better than it. However, it can be observed from the visualization results that the addition strategy will enhance the salient features too much and form over-exposed images. Although this strategy improves the metrics from introducing too much information, our method can achieve better visual effects, which is why we need to choose multiple metrics from different dimensions to evaluate the quality of the generated images.

V. CONCLUSION

In this study, we propose a novel fusion network, termed DePF (Decomposition Pooling based Fusion), designed for the fusion of infrared and visible images. DePF comprises three key components: an encoder, a fusion layer, and a decoder. Notably, our approach diverges from conventional wavelet-based pooling in de-pooling, instead prioritizing an expanded receptive field to extract more extensive deep information. Furthermore, we modify the feature extraction kernel based on the Prewitt operation, enabling the extraction of multi-scale information while simultaneously preserving fine-grained details. It is worth highlighting that our de-unpooling process ensures the precise recovery, thereby enabling our model to effectively retain structural intensities, conserve intricate details, and reconstruct images faithfully. Additionally, our fusion network incorporates a spatial attention fusion strategy, which is both straightforward and efficient.

To assess the efficacy of our proposed method, we conduct experiments using three publicly available datasets. Our approach consistently outperforms nine state-of-the-art methods in terms of both visualization quality and quantitative evaluation. Furthermore, extensive experimentation demonstrates that our method excels in processing efficiency, object detection, and noise-related evaluations. Ablation experiments corroborate the effectiveness of individual components within our proposed method.

Nevertheless, our DePF currently limits its capacity to extract multi-scale information and intricate textures to a

predefined set of four directions. In subsequent research endeavors, we intend to devise innovative pooling techniques to delve into the extraction of multi-directional feature information or the autonomous acquisition of features spanning diverse directions, thereby addressing a broader spectrum of challenges.

The fusion of infrared and visible images holds paramount significance in diverse fields, as it enables to enhance visual perception and analytical capabilities. This multidisciplinary technique finds applications in areas such as surveillance, detection, segmentation and more.

REFERENCES

[1] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[2] X. Chang, L. Jiao, F. Liu, and F. Xin, "Multicontourlet-based adaptive fusion of infrared and visible remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 3, pp. 549–553, 2010.

[3] Y. Liu, L. Dong, and W. Xu, "Infrared and visible image fusion for shipborne electro-optical pod in maritime environment," *Infrared Physics & Technology*, vol. 128, p. 104526, 2023.

[4] N. Paramanandham and K. Rajendiran, "Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications," *Infrared Physics & Technology*, vol. 88, pp. 13–22, 2018.

[5] D. Li, Z. Song, C. Quan, X. Xu, and C. Liu, "Recent advances in image fusion technology in agriculture," *Computers and Electronics in Agriculture*, vol. 191, p. 106491, 2021.

[6] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, "Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4224–4231, 2018.

[7] R. Guo, D. Li, and Y. Han, "Deep multi-scale and multi-modal fusion for 3d object detection," *Pattern Recognition Letters*, vol. 151, pp. 236–242, 2021.

[8] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2091–2106, 2021.

[9] D. Wang, J. Liu, R. Liu, and X. Fan, "An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection," *Information Fusion*, vol. 98, p. 101828, 2023.

[10] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. Van De Weijer, and F. Shahbaz Khan, "Multi-modal fusion for end-to-end rgb-t tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[11] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9516–9526.

[12] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Information fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[13] 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," *Information fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[14] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar conference on signals, systems and computers*. IEEE, 1993, pp. 40–44.

[15] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Transactions on biomedical engineering*, vol. 59, no. 12, pp. 3450–3459, 2012.

[16] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2760–2765.

[17] B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," *Information fusion*, vol. 13, no. 1, pp. 10–19, 2012.

[18] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 2018, pp. 2705–2710.

[19] C. Gao, C. Song, Y. Zhang, D. Qi, and Y. Yu, "Improving the performance of infrared and visible image fusion based on latent low-rank representation nested with rolling guided image filtering," *IEEE Access*, vol. 9, pp. 91 462–91 475, 2021.

[20] D. Bhavana, K. Kishore Kumar, and D. Ravi Tej, "Infrared and visible image fusion using latent low rank technique for surveillance applications," *International Journal of Speech Technology*, pp. 1–10, 2021.

[21] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[22] 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 Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9645–9656, 2020.

[23] H. Li, Y. Xiao, C. Cheng, and X. Song, "Sfpfusion: An improved vision transformer combining super feature attention and wavelet-guided pooling for infrared and visible images fusion," *Sensors*, vol. 23, no. 18, p. 7870, 2023.

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

[25] Y. Xiao, H. Li, C. Cheng, and X. Song, "Le2fusion: A novel local edge enhancement module for infrared and visible image fusion," *arXiv preprint arXiv:2305.17374*, 2023.

[26] H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler, "LRRNet: A Novel Representation Learning Guided Fusion Network for Infrared and Visible Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[28] 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," *Applied Soft Computing*, vol. 132, p. 109857, 2023.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014, p. 2672–2680.

[30] 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," *Information Fusion*, vol. 82, pp. 28–42, 2022.

[31] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "Divfusion: Darkness-free infrared and visible image fusion," *Information Fusion*, vol. 91, pp. 477–493, 2023.

[32] C. Cheng, T. Xu, and X.-J. Wu, "Mufusion: A general unsupervised image fusion network based on memory unit," *Information Fusion*, vol. 92, pp. 80–92, 2023.

[33] J. Liu, R. Dian, S. Li, and H. Liu, "Sgfusion: A saliency guided deep-learning framework for pixel-level image fusion," *Information Fusion*, vol. 91, pp. 205–214, 2023.

[34] R. Dian, A. Guo, and S. Li, "Zero-shot hyperspectral sharpening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[35] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5618–5627.

[36] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," *arXiv preprint arXiv:1805.08620*, 2018.

[37] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *International conference on learning representations*, 2018.

[38] B.-S. Wang, J.-W. Hsieh, P.-Y. Chen, M.-C. Chang, L. Ke, and S. Lyu, "Ldw-pooling: Learnable discrete wavelet pooling for convolutional networks," 2021.

[39] G. Pajares and J. M. De La Cruz, "A wavelet-based image fusion tutorial," *Pattern recognition*, vol. 37, no. 9, pp. 1855–1872, 2004.

[40] M. N. Do and M. Vetterli, "Contourlets: a directional multiresolution image representation," in *Proceedings. International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. I–357–I–360.

[41] G. Easley, D. Labate, and W.-Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 25–46, 2008.

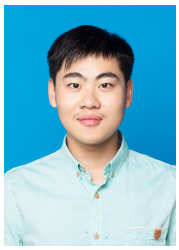
[42] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Processing*, vol. 9, no. 5, pp. 347–357, 2015.

[43] B. Yang and S. Li, "Visual attention guided image fusion with sparse representation," *Optik*, vol. 125, no. 17, pp. 4881–4888, 2014.

[44] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.
[45] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4714–4722.
[46] L. Qu, S. Liu, M. Wang, and Z. Song, "Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2126–2134.
[47] 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 Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2020.
[48] 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 Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1186–1196, 2021.
[49] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "Ifcnn: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.
[50] M. Xu, L. Tang, H. Zhang, and J. Ma, "Infrared and visible image fusion via parallel scene and texture learning," *Pattern Recognition*, vol. 132, p. 108929, 2022.
[51] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, vol. 83, pp. 79–92, 2022.
[52] X. Zhang, H. Zhai, J. Liu, Z. Wang, and H. Sun, "Real-time infrared and visible image fusion network using adaptive pixel weighting strategy," *Information Fusion*, p. 101863, 2023.
[53] Y. Rao, D. Wu, M. Han, T. Wang, Y. Yang, T. Lei, C. Zhou, H. Bai, and L. Xing, "At-gan: A generative adversarial network with attention and transition for infrared and visible image fusion," *Information Fusion*, vol. 92, pp. 336–349, 2023.
[54] D. Rao, T. Xu, and X.-J. Wu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Transactions on Image Processing*, 2023.
[55] 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 Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.
[56] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "Swinfuse: A residual swin transformer fusion network for infrared and visible images," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.
[57] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "Datfuse: Infrared and visible image fusion via dual attention transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
[58] T. Chen and J. Yang, "Very deep fully convolutional encoder–decoder network based on wavelet transform for art image fusion in cloud computing environment," *Evolving Systems*, vol. 14, no. 2, pp. 281–293, 2023.
[59] Z. Wang, X. Li, H. Duan, X. Zhang, and H. Wang, "Multifocus image fusion using convolutional neural networks in the discrete wavelet transform domain," *Multimedia Tools and Applications*, vol. 78, pp. 34 483–34 512, 2019.
[60] R. Xu, G. Liu, Y. Xie, B. D. Prasad, Y. Qian, and M. Xing, "Multiscale feature pyramid network based on activity level weight selection for infrared and visible image fusion," *JOSA A*, vol. 39, no. 12, pp. 2193–2204, 2022.
[61] Z. Chao, X. Duan, S. Jia, X. Guo, H. Liu, and F. Jia, "Medical image fusion via discrete stationary wavelet transform and an enhanced radial basis function neural network," *Applied Soft Computing*, vol. 118, p. 108542, 2022.
[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[63] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
[64] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3496–3504.
[65] A. Toet, "The tno multiband image data collection," *Data in brief*, vol. 15, pp. 249–251, 2017.
[66] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2020.
[67] 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," *Computer Vision and Image Understanding*, vol. 218, p. 103407, 2022.
[68] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
[69] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.
[70] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9036–9045.

Hui Li received the B.Sc. degree in School of Internet of Things Engineering from Jiangnan University, China, in 2015. He received the PhD degree at the School of Internet of Things Engineering, Jiangnan University, Wuxi, China, in 2022. He is currently a Lecturer at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include image fusion and multi-modal visual information processing. He has been chosen among the World's Top 2% Scientists ranking in the single recent year dataset published by Stanford University (2022 and 2023).

He has published several scientific papers, including IEEE TPAMI, IEEE TIP, Information Fusion, IEEE TCYB, IEEE TIM, ICPR etc. He achieved top tracking performance in several competitions, including the VOT2020 RGBT challenge (ECCV20) and Anti-UAV challenge (ICCV21).



Yongbiao Xiao received the B.Sc. degree in computer science and technology from Suzhou University of Science and Technology, Suzhou, China, in 2022. He is currently pursuing the M.S. degree with International Joint Laboratory on Artificial Intelligence of Jiangsu Province, Jiangnan University, Wuxi, China.

His research interests include image fusion and deep learning.

Chunyang Cheng is working toward the Ph.D degree at Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, China. His research interests include image fusion and deep learning.

Zhongwei Shen received the M.S. degree in image processing and multi-media from INP-ENSEEIHT, Toulouse, France, in 2015. He received the PhD degree at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China, in 2022. He is currently a lecturer at the School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China. His research interests include video analysis, action recognition and deep learning. He has published several scientific papers, including IEEE TMM, IEEE TCSVT etc.

Xiaoning Song received the BSc degree in computer science from Southeast University, Nanjing, China, in 1997, the MSc degree in computer science from the Jiangsu University of Science and Technology, Zhenjiang, China, in 2005, and the PhD degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, in 2010. He was a visiting researcher with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, UK, from 2014 to 2015. He is currently a full Professor with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include pattern recognition, machine learning, and computer vision.