Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 33(6): 4045–4073.
DOI: 10.3934/era.2025180
Received: 17 April 2025
Revised: 9 June 2025
Accepted: 13 June 2025
Published: 27 June 2025

Research article

TSTBFuse: a two-stage three-branch feature extraction method for infrared and visible image fusion

Wangwei Zhang[1], Xinyue Qin[1], Menghao Dai[1], Bin Zhou[2,*], Changhai Wang[1], ZhiHeng Wang[3], SongZe Li[4]

[1] Software Engineering College, Zhengzhou University of Light Industry, No.136 Science Road, Zhengzhou 450000, China

[2] Electronics and Electrical Engineering College, Zhengzhou University of Science and Technology, No.1 Xueyuan Road, Zhengzhou 450064, China

[3] Zhengzhou Xinda Institute of Advanced Technology, Zhengzhou, China

[4] Henan Xindawangyu Science & Technology Co., Ltd., Zhengzhou, China

* **Correspondence:** Email: whelmmail@126.com; Tel: +8617539126677.

Abstract: The purpose of image fusion is to combine information from different source images to produce a comprehensively representative image. Traditional autoencoder architectures often struggle to effectively extract both unique and shared features from these image types. A novel two-stage three-branch feature extraction method (TSTBFuse) was proposed in the study, specialized for the fusion of infrared and visible images. The proposed architecture employed a three-branch encoder that separately captured infrared-specific thermal radiation features, visible-specific texture details, and shared structural information. A two-stage end-to-end training strategy was introduced: the first stage focused on reconstructing the original input images to preserve modality-specific information, while the second stage leveraged the learned representations to generate high-quality fused images. we designed a comprehensive loss function combining mean squared error (MSE), structural similarity index (SSIM), and gradient loss, ensuring both pixel-level accuracy and structural integrity. Extensive experiments on public datasets (TNO, MSRS and RoadScene) demonstrated that TSTBFuse consistently outperformed seven state-of-the-art methods in both subjective and objective evaluations. Furthermore, the method exhibited strong generalization capabilities, successfully extending to challenging tasks such as magnetic resonance imaging-computed tomography (MRI-CT) medical image fusion and red-green-blue (RGB)-infrared image fusion without retraining. The code is publicly available at: https://github.com/QXinYue/TSTBFuse.

Keywords: image fusion; shared features; three-branch feature extraction; convolutional neural network

1. Introduction

With the rapid advancement of sensor technology, the field of image fusion has garnered significant attention in recent research [1]. The primary objective of image fusion is to integrate information from diverse modalities while eliminating redundant components, thereby enhancing the expressive capability and practical applicability of the resulting images. Infrared images are effective at capturing thermal target characteristics by measuring the thermal radiation emitted by objects; however, they exhibit low sensitivity to variations in scene brightness and often lack clarity. In contrast, visible images provide rich scene details and high definition, although their ability to convey target features is relatively limited. By fusing infrared and visible images, composite images can be generated that exhibit both high-contrast target features and detailed scene information, thereby improving the overall comprehensiveness of the data. Such fused images hold promising potential for applications in areas such as intelligent driving [2], semantic segmentation [3], target detection [4], and recognition [5, 6].

In this context, the rapid development of deep learning has significantly evolved methods for fusing infrared and visible images. Early pixel-level fusion methods were predominantly based on straightforward techniques, such as weighted averaging or pixel value superposition for image synthesis [7–9]. Feature-based approaches, including scale-Invariant feature transform (SIFT) and speeded up robust features (SURF), focused on extracting structural information from the images [10]. Transform-domain fusion methods involved converting images into alternative domains, such as wavelet transforms, discrete cosine transforms, or Laplacian pyramids, to extract meaningful features for fusion [11–14]. After feature extraction, an inverse transform was applied to reconstruct the fused image in the spatial domain. While transform-domain approaches have proven effective in preserving fine image details and features, they are often computationally intensive. Despite the advancements achieved by early methods, challenges persist, particularly in maintaining both target and fine detail features when fusing images from different modalities.

In recent years, numerous deep learning-based methods have been developed for fusing infrared and visible images. For instance, convolutional neural networks (CNNs) have been utilized to enhance fusion by inputting source images from different modalities for feature extraction and learning fusion strategies [15–17]. Generative adversarial networks (GANs) have been employed to generate high-quality fused images through the interaction between a generator and a discriminator, where the generator produces images and the discriminator evaluates the differences between the generated and real fused images [18, 19]. Additionally, autoencoder (AE) approaches have been used to extract features via an encoder and reconstruct the fused image using a decoder [16, 20, 21]. However, CNN-based methods are limited by their relatively static feature extraction process, which restricts their ability to capture complex relationships between modalities. Although GANs can generate high-quality images, they suffer from unstable training processes. The adversarial nature of GANs can lead to issues such as training failures, mode collapse, and the presence of artifacts in fused images. To address these challenges, we propose a novel two-stage three-branch feature extraction method (TSTBFuse) for infrared and visible image fusion, aiming to improve fusion quality.

In the contemporary scientific landscape, AE methods have gained significant traction. However, many AE approaches face challenges in effectively extracting shared features between infrared and visible images, as well as capturing the fine texture details of infrared images and the thermal radiation characteristics of visible images during the encoding phase. Current AE architectures can be

classified into three primary forms, as depicted in Figure 1(a)–(c). As shown in Figure 1(a), the method excels at extracting shared features but struggles to capture the distinctive features of both image types. Figure 1(b) illustrates an approach in which two independent encoders effectively capture unique features but largely overlook shared ones. Meanwhile, Figure 1(c) employs a shared encoder for feature extraction, followed by independent encoders that capture both base and detailed features. While this method ultimately synthesizes the features and reconstructs the image through a decoder, it still results in a significant loss of shared features. These limitations highlight a fundamental trade-off in existing AE-based fusion methods: they typically prioritize unique features or shared features, but rarely consider both simultaneously. To address these challenges, we propose a novel AE architecture, TSTBFuse, as shown in Figure 1(d). During the encoding phase, infrared and visible light images are input into the same network via channel concatenation, and two independent encoder networks are used to capture the unique features of each image type separately. In essence, TSTBFuse explicitly separates the extraction of modality-specific features from shared structural features. This separation ensures that unique and shared information is preserved without interference, enabling more effective multimodal fusion. To further optimize the feature extraction process, a two-stage training strategy is proposed. During the decoding phase, the TSTBFuse method integrates the Restormer module with a CNN, enhancing image reconstruction and providing a more balanced and effective solution to the challenges of feature extraction and image synthesis.

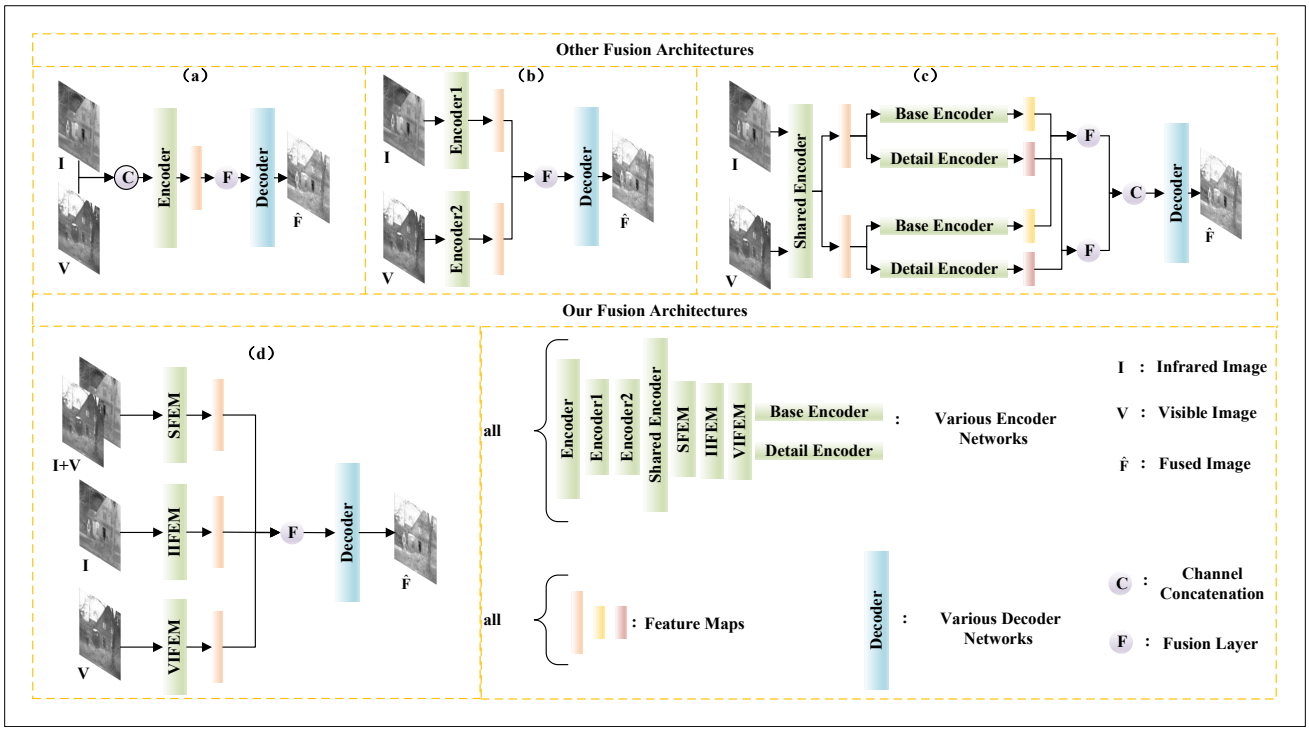

Figure 1. (a), (b), and (c) depict various fused self-encoder structures, while (d) illustrates the TSTBFuse architecture. In this architecture, IIFEM, VIFEM, and SFEM represent the infrared feature extraction module, the visible feature extraction module, and the shared feature extraction module, respectively.

To overcome the limitations of existing self-encoder methods in extracting both shared and unique

features from infrared and visible images, as well as to address deficiencies in fusion strategies and image reconstruction, the proposed TSTBFuse fusion method introduces four key innovative contributions:

- In this study a three-branch feature extraction method is proposed to tackle the challenge of independently extracting shared and unique features. This method incorporates an infrared image feature extraction module, a visible image feature extraction module, and a shared feature extraction module, enabling the effective extraction of both shared and unique features from infrared and visible images.
- This study proposes a two-stage end-to-end training strategy for the TSTBFuse method. In the first stage, shared and unique features from infrared and visible images are combined through channel splicing to efficiently recover the original image. In the second stage, these features are fully fed into the decoder network, resulting in a high-quality fused image reconstruction. This approach eliminates intermediate steps and manual feature extraction, significantly improving both the model's efficiency and the quality of image fusion.
- This study proposes a triple loss fusion strategy, introducing a novel integrated loss function to enhance the reconstruction of infrared and visible images. The function combines MSE, SSIM and gradient loss, effectively capturing pixel-level differences while enhancing structural information and detail clarity. The integrated loss function ensures that the final fused image retains the features of both image modalities, resulting in high-quality image restoration and significant improvements in visual quality.
- Compared to existing state-of-the-art fusion methods, TSTBFuse demonstrates significant performance improvements in both subjective and objective evaluations across three public datasets: thermal and near-infrared optical (TNO), multispectral stereo registration scene (MSRS), and RoadScene. Additionally, excellent results are achieved in subsequent tasks, including RGB fusion tests and medical image datasets such as MRI-CT, further validating its effectiveness and innovative contributions to the field of image fusion.

2. Related works

In Sections 2.1 and 2.2, this paper presents traditional infrared and visible image fusion algorithms, along with related methods based on deep learning. Section 2.3 will then focus on the study of a three-branch encoder architecture.

2.1. Traditional infrared and visible image fusion algorithms

Traditional fusion of infrared and visible images primarily relies on mathematical transformations to establish fusion rules by analyzing the spatial or transform domains. Common techniques include methods based on sparse representation, as well as those operating in the spatial and transform domains.

The sparse representation-based approach begins by vectorizing the infrared and visible images into pixel vectors, which are then converted into sparse coefficients using a sparse coding mechanism and the sparsity of an overcomplete dictionary. These sparse coefficients are processed using strategies such as maximum sparse selection, average fusion, weighted fusion, and region selection to achieve

region-specific fusion. Finally, the fused image is reconstructed through inverse transformation. This method heavily depends on dictionary learning, and poor dictionary selection can result in suboptimal reconstruction, leading to the loss of image details. In recent years, the integration of this approach with deep learning has significantly enhanced its performance. For example, Li et al. [22] proposed an adaptive dictionary learning framework that improved the fusion effect; Zhang et al. [23] employed a multi-scale sparse representation method to enhance detail and contrast; Wang et al. [24] refined the sparse coding algorithm by incorporating deep learning techniques; and Huang et al. [25] developed a sparse representation fusion model based on deep learning, which improved performance in processing complex scenes.

Spatial domain-based fusion methods primarily operate on the pixel domain of source images and include techniques such as weighted averaging, maximum value selection, and multi-scale fusion. These methods often necessitate the design of specific rules tailored to particular requirements. For instance, Gonzalez et al. [26] proposed a weighted average method for effective image fusion; Pohl et al. [27] examined the concept and application of multi-sensor image fusion using maximum value selection; Burt et al. [28] introduced a multi-scale fusion technique based on Laplace pyramid compact coding; Chakrabarti et al. [29] utilized logarithmic transformation to analyze the statistical features of real-world illumination; and Salgado et al. [30] investigated the application of image enhancement techniques in endoscopic systems to improve image quality.

Transform domain-based fusion methods extract features by converting infrared or visible images into different domains to identify their similarities and differences. These images are then processed using appropriate fusion strategies, with the final fused image obtained through inverse transformation. For example, discrete wavelet transform (DWT) is widely employed for image fusion. Tian et al. [31] effectively used wavelet transform to extract multi-scale features and enhance image details, while Gao et al. [32] applied Fourier Transform to improve image contrast. However, relying solely on a single wavelet or Fourier transform may result in a loss of spatial information, adversely affecting the quality of the fused image. To address this issue, Xiong et al. [15] proposed a method that combines wavelet transform and principal component analysis, enabling the retention of both spatial and frequency domain information to enhance image fusion.

2.2. Deep learning-based fusion method for infrared and visible images

Deep learning-based fusion methods for infrared and visible images primarily include approaches based on CNNs, GANs, and AE architectures. These methods focus on addressing three key challenges: feature extraction, feature fusion, and image reconstruction.

CNN-based fusion methods primarily enhance network performance by refining the loss function and fusion strategy. For instance, Li et al. [33] proposed a deep prediction network, IVFuseNet, which adaptively weights fusion based on the information content of the two images. This network automatically adjusts the generated image under the constraints of a loss function, reducing the need for manual intervention in fusion strategy selection. However, the network's relatively simple structure limits its ability to ensure high-quality fusion. Zhang et al. [34] introduced a generalized image fusion framework, Image Fusion Convolutional Neural Network (IFCNN), which demonstrates good versatility but struggles with detail preservation and is prone to artifacts.

GAN-based fusion methods comprise two primary components: a generator and a discriminator. The generator creates fused images, while the discriminator assesses whether an input image is real or

synthesized. When the discriminator identifies an image as fake, the generator adjusts its parameters based on the discriminator's feedback, iteratively training to produce the final fused image. Ma et al. [18] was the first to apply GANs to infrared and visible image fusion, proposing FusionGAN, which generates fused images through adversarial training. However, FusionGAN exhibits limitations in preserving thermal radiation information from infrared images and fine details from visible images. To mitigate these issues, researchers introduced loss functions such as detail loss and target edge enhancement loss to better preserve image details and sharpen edges. Yue et al. [35] later proposed DifFusion, an image fusion network based on the diffusion model, which significantly enhances multi-source information integration and improves color fidelity.

Fusion methods based on self-encoder architectures typically comprise three components: an encoder, a fusion layer, and a decoder. This approach extracts features via the encoder, applies a fusion strategy, and inputs the fused features processed by the fusion layer into the decoder to generate the final fused image. For instance, Huang et al. [36] proposed DenseNet, a residual network architecture that effectively preserves source image features through interconnections between feature layers. Li et al. [16] extended this work with DenseFuse, which employs densely connected convolutional layers to extract more features from the source image and offers two fusion strategies: weighting and averaging. To further enhance feature extraction, Xu et al. [37] introduced common and unique feature eecomposition (CUFD), a dual encoder-decoder network that efficiently captures both deep and shallow features, fusing them using weighted averaging and maximal blending rules. Furthermore, Tang et al. [38] addressed the limitations in extracting global contextual information by incorporating a Transformer module and designing a dual-attention fusion method, DATFuse, which effectively models global contextual dependencies to produce fusion results with enhanced texture details.

2.3. Study of the three-branch encoder structure

Recent studies have explored various network designs for infrared and visible image fusion, with particular emphasis on dual-branch architectures and attention mechanisms. For instance, hybrid CNN-Transformer networks for medical image fusion [43], detail-semantic-aware fusion (DSAFusion) [39] employs a dual-branch encoder to extract detail and semantic features separately, followed by dedicated fusion networks. Target-aware Taylor expansion approximation (T2EA) [40] introduces a Taylor expansion decomposition strategy and a dual-branch fusion design to highlight target information. Multi-stage feature learning with channel-spatial attention (MSCS) [41] proposes a multi-stage feature learning approach with channel-spatial attention to handle illumination variations, while visible-infrared fusion network (VIF-Net) [42] leverages an unsupervised learning framework based on a mixed loss function to enhance thermal and texture detail fusion.

However, these methods share common limitations: they often focus on either modality-specific features or shared features, but rarely both; they lack explicit disentanglement of feature types, leading to information redundancy or loss of structural details; and they may struggle to adapt across diverse scenarios, such as varying illumination or complex scenes. Furthermore, most existing methods lack a progressive learning strategy that systematically refines feature extraction and fusion performance.

To address these gaps, TSTBFuse introduces a dedicated three-branch encoder structure that explicitly separates infrared-specific, visible-specific, and shared features, reducing cross-modality interference. Additionally, a two-stage training strategy is designed to progressively optimize feature

learning and reconstruction quality, enabling TSTBFuse to achieve a more balanced, interpretable, and generalizable fusion result compared to existing approaches. In this article, the objective is to extract infrared features, visible light features, and their shared features more efficiently. To this end, three specialized network modules have been designed: IIFEM, VIFEM, and SFEM.

The IIFEM is designed based on the principles of the Inception module [45] and aims to improve the efficiency of thermal radiation feature extraction from infrared images. This module employs five branches for multi-scale feature extraction. Initially, the infrared images are processed through a 1×1 convolutional layer, with the number of channels adjusted to 64 via mean normalization (MN layer) and rectified linear unit (ReLU) activation. Subsequently, the features pass through four distinct branches consisting of 1×1, 3×3, and 7×7 convolutional layers, as well as a 3×3 max-pooling layer, enabling the extraction of features at multiple scales. The outputs of these branches undergo mean normalization and ReLU activation, yielding feature maps with 63, 64, 64, and 64 channels, respectively.

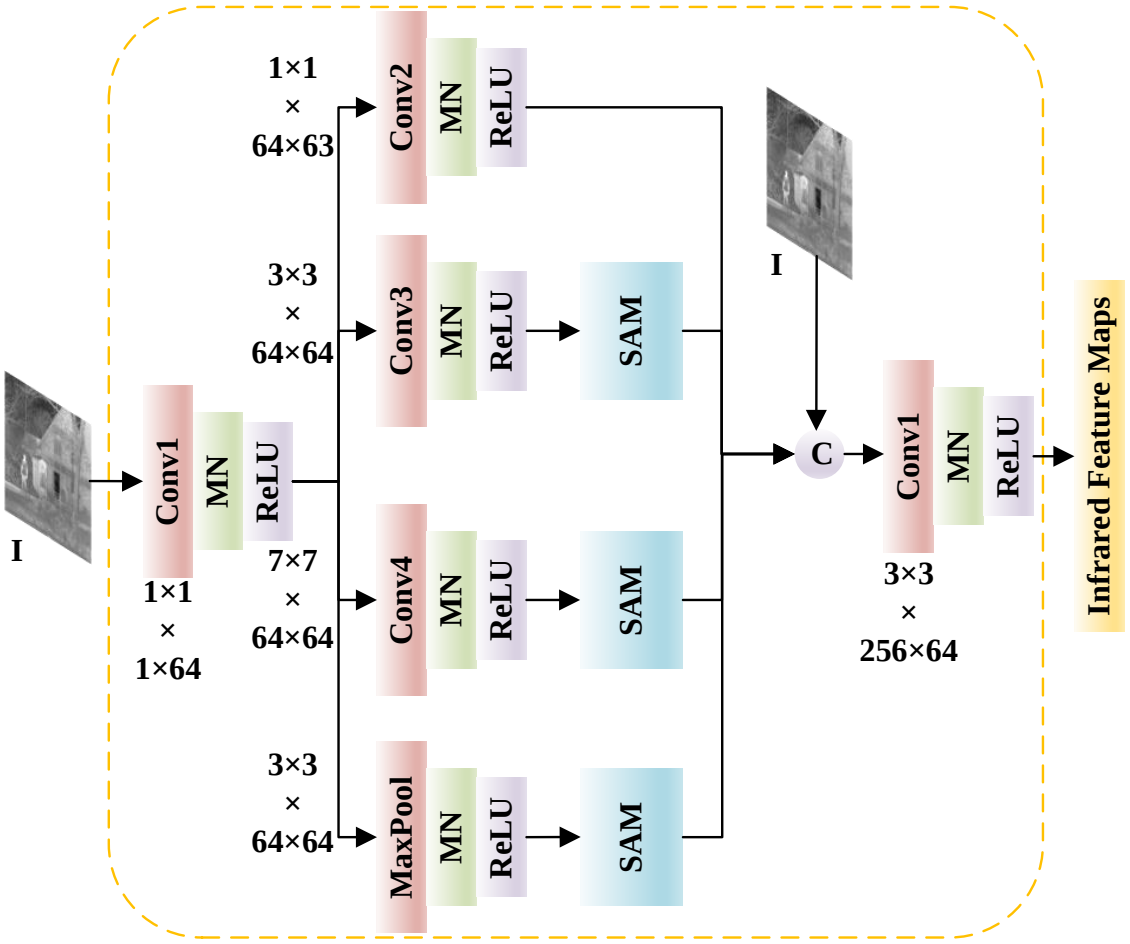

Figure 2. Infrared feature extraction module.

To emphasize salient features and enhance feature representation, a spatial attention mechanism is incorporated after the Conv3 and Conv4 layers, further improving the module's ability to capture

critical information. Additionally, to preserve details of the infrared image and enrich semantic information, the original infrared image is channel-wise concatenated with the four convolved feature maps, resulting in a 256-channel input feature layer. These features are then processed through a 3 × 3 convolutional layer to generate the optimal feature map for the infrared image. This design not only boosts feature extraction effectiveness but also enhances the model's adaptability to complex scenes. The architecture of the infrared feature extraction module is illustrated in Figure 2.

To efficiently extract detailed features from visible images in both channel and spatial dimensions, the module is designed with two parallel branches for feature extraction. Initially, the visible light image undergoes feature reorganization via a 3 × 3 convolutional layer, with the number of channels adjusted to 32 using mean normalization (MN layer) and ReLU activation. Subsequently, the resulting feature maps are fed into a spatial attention module (SAM) through one branch, allowing the network to automatically identify and emphasize the most relevant spatial features. Concurrently, the other branch directs the convolved feature maps into a channel attention module (CAM), which dynamically focuses on critical channels within the visible image, thereby enhancing the capture of key spatial information. Finally, the 64-channel feature maps processed by SAM and CAM are concatenated along the channel dimension to produce the final 64-channel visible image features. This entire system is referred to as the visible image feature extraction module (VIFEM), as illustrated in Figure 3.

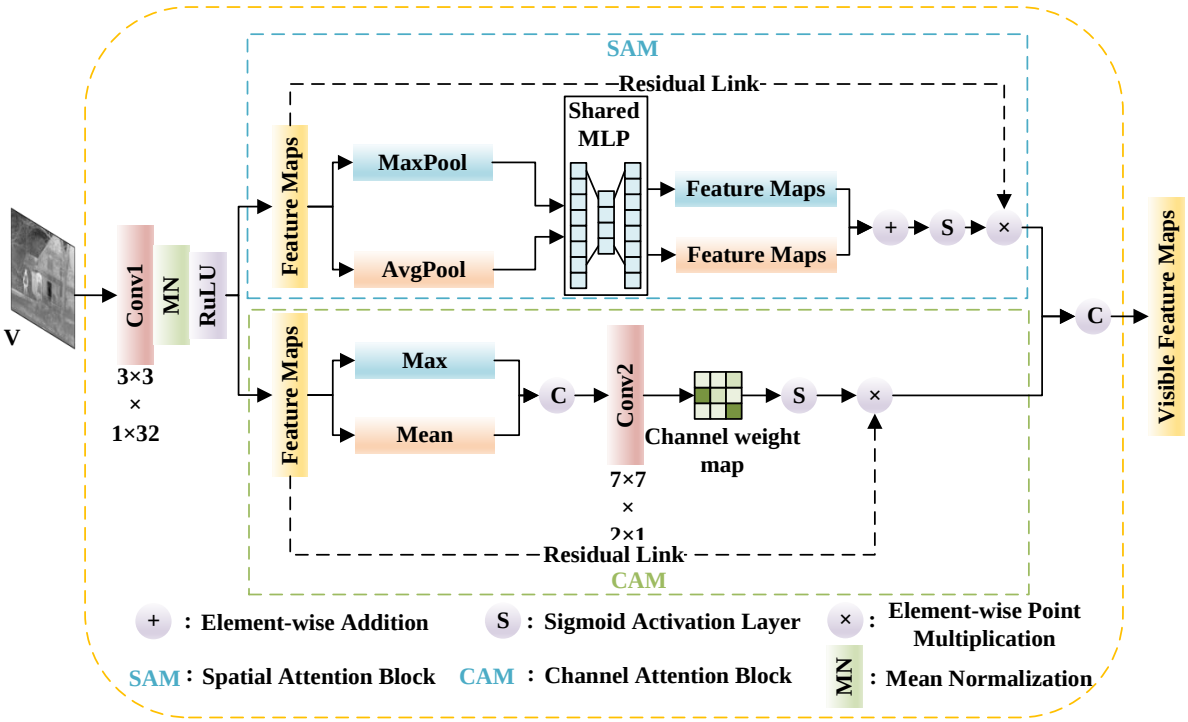

Figure 3. Visible image feature extraction module.

The SFEM is inspired by the DenseBlock architecture proposed by Li et al. This module is structured to ensure that the inputs and outputs of each layer conform to the design principles of the DenseBlock module, thereby facilitating the efficient extraction of shared features from both infrared and visible light images. Initially, the infrared and visible images are concatenated along the channel

dimension and fed into the first convolutional layer. Subsequently, the output of each layer serves as input to all succeeding layers. After five layers are densely concatenated, a shared feature map with 64 channels is obtained. Based on this design, the module is designated as the SFEM, as illustrated in Figure 4 below.

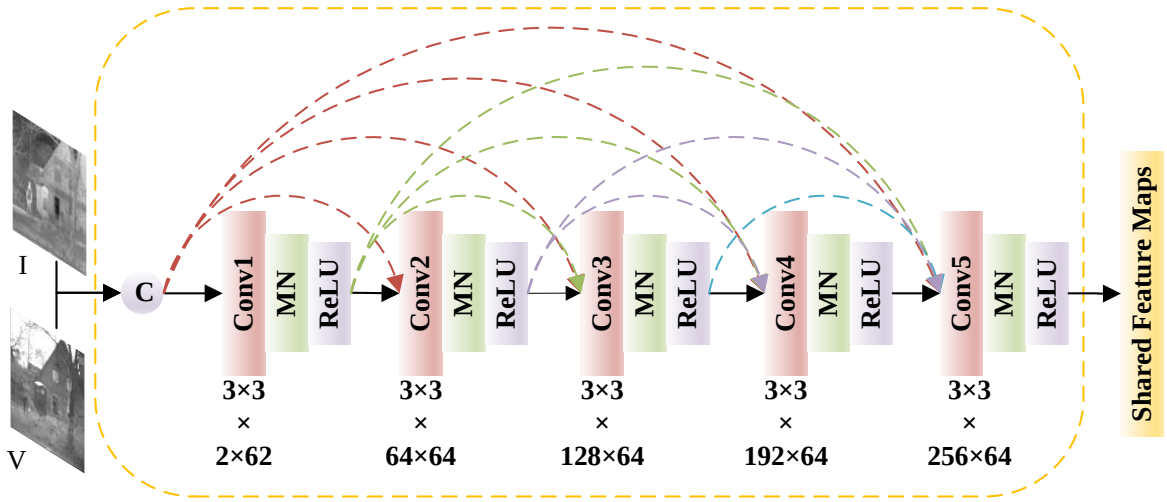

Figure 4. Shared feature extraction module.

3. Methodology

This section first introduces the workflow of TSTBFuse and the detailed architecture of its constituent modules, followed by a discussion on the design of the two-stage fusion strategy and the corresponding loss function.

3.1. Overview

In multi-modal fusion tasks such as infrared and visible image fusion, it is essential to effectively extract and preserve both modality-specific features and shared features across different domains. The proposed three-branch encoder architecture of TSTBFuse is theoretically motivated by the principles of shared-private representation learning and multitask learning theory. TSTBFuse is a method based on a self-encoder architecture, primarily consisting of an encoder, a two-stage fusion strategy, and a decoder. In the first stage of training, emphasis is placed on recovering the original image, while in the second stage, the focus is on generating the fused image. The workflow of TSTBFuse is illustrated in detail in Figure 5.

The model consists of approximately 2 million parameters and achieves an average inference time of 22 milliseconds per 256×256 images on an RTX 4090 GPU. The memory usage during the inference process is approximately 520 MB, indicating that despite the use of multi-branch and transformer-based components, resource requirements are still relatively low. Compared with existing methods, our approach achieves a balance between fusion quality and efficiency, and future work can further

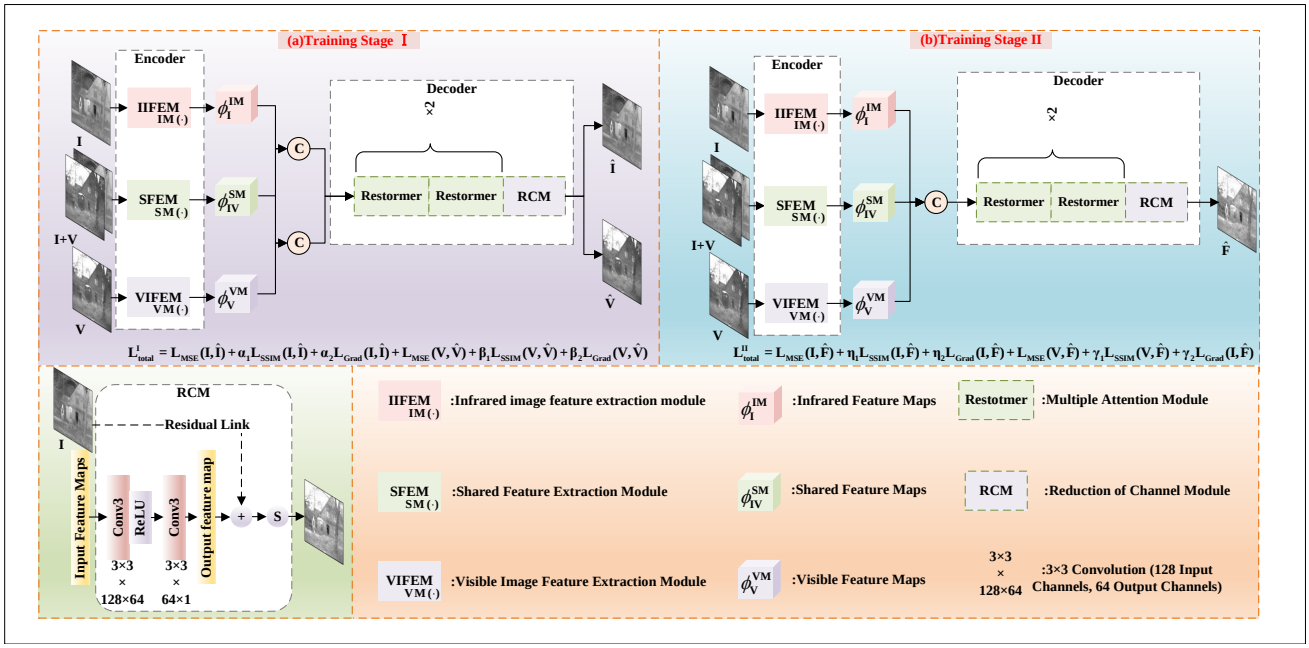

Figure 5. Overall framework diagram of the TSTBFuse method. The inputs include the infrared image (I), visible image (V), and the channel-spliced images of I and V.

improve performance through model pruning or quantization.Although TSTBFuse uses a three-branch encoder and a Transformer-based decoder, its total number of parameters is approximately 2 million, making it a lightweight model comparable in scale to methods such as IFCNN and DenseFuse. It offers a good balance between performance and complexity.

3.2. Encoders

The encoder structure comprises three branches: the IIFEM, the VIFEM, and the SFEM, each of which has been detailed in the previous section. To further clarify the workflow of TSTBFuse, we introduce the following notations. The input single-channel infrared and visible images are represented by $I \in R^{H*W}$ and $V \in R^{H*W}$, respectively. The functions of IIFEM, VIFEM, and SFEM are denoted as $IM(\cdot)$, $VM(\cdot)$, and $SM(\cdot)$. The process of channel stitching is indicated by the following.

First, to more effectively extract unique features from infrared images, we utilize the infrared image feature extraction module (IIFEM). These unique features are denoted by the symbol ϕ_I^{IM}, i.e.,

$$\Phi_I^{IM} = IM(I) \tag{3.1}$$

Meanwhile, the visible image feature extraction module (VIFEM) is used to extract the unique features of the visible image, denoted as ϕ_V^{VM}, i.e.,

$$\Phi_V^{VM} = VM(V) \tag{3.2}$$

In addition, the shared feature extraction module (SFEM) is used to extract shared features of

infrared and visible images, denoted as ϕ_{IV}^{SM}, i.e.,

$$\Phi_{IV}^{SM} = SM(C(I, V)) \tag{3.3}$$

By extracting features through these three modules, both the unique and shared characteristics of infrared and visible images can be comprehensively captured, thereby establishing a solid foundation for the subsequent fusion process. To improve fusion output quality, we adopt the Restormer module as the core architecture of the decoder. Restormer is a Transformer-based image restoration framework designed to efficiently model long-range dependencies. It introduces two key components: multi-scale convolutional head transposed attention (MDTA), which captures spatial relationships between distant regions, and a gated deep feedforward network that adaptively refines local features. In our method, the Restormer module helps the decoder better integrate multiple feature streams by simultaneously capturing global context and fine-grained details, thereby generating clearer and more natural fusion outputs. Compared to decoders that use only CNNs, Restormer demonstrates superior performance in terms of structural preservation and texture clarity, particularly in complex fusion scenarios.

3.3. Design of a two-stage integration strategy

TSTBFuse employs an end-to-end training scheme to implement the fusion strategy via channel splicing. During the first stage of training, in order to retain the original image information, the features ϕ_I^{IM}, ϕ_V^{VM}, which are specific to the infrared or visible images and extracted by the three-branch encoder, are each channel spliced with the shared features ϕ_{IV}^{SM}. This process generates the infrared fusion layer ϕ^{IS} and the visible fusion layer ϕ^{VS}, both incorporating the shared features, i.e.,

$$\Phi^{IS} = C(\phi_I^{IM}, \phi_{IV}^{SM}) \tag{3.4}$$

$$\Phi^{VS} = C(\phi_V^{VM}, \phi_{IV}^{SM}) \tag{3.5}$$

In the second training stage, to generate the fused image, the infrared and visible features are simultaneously concatenated with the shared features along the channel dimension, producing a feature map A that contains both unique and shared features, i.e.,

$$\Phi^{IVS} = C(\phi_I^{IM}, \phi_V^{VM}, \phi_{IV}^{SM}) \tag{3.6}$$

The two-stage training strategy plays a key role in decoupling and preserving complementary modal information. In the first stage, infrared and visible light image features are connected with shared features along the channel dimension, enabling each branch to retain modality-specific information, with the aim of improving the decoder's ability to reconstruct the original image. In the second stage, the parameters learned in the first stage are retained to more effectively utilize information from infrared features, visible light features, and their shared representations. This enables the decoder to integrate well-separated and semantically rich features, achieving a better balance between structural fidelity and texture enhancement. Stage-wise optimization thus helps avoid premature feature mixing, enhances cross-modal complementarity, and produces higher-quality fusion outputs.

3.4. Loss function design

During the two-stage training process, a combined loss function comprising MSE loss, SSIM loss, and gradient loss is employed in the first stage to more effectively reconstruct the original infrared and visible images. The MSE loss primarily quantifies pixel-level differences between the reconstructed and ground-truth images, efficiently capturing global errors and guiding the model to enhance fine detail recovery. Conversely, the SSIM loss emphasizes the preservation of structural information, ensuring that the reconstructed image is perceptually similar to the original by assessing luminance, contrast, and structural components. The combination of MSE and SSIM significantly enhances the visual quality of the images. Furthermore, the gradient loss refines edge sharpness and detail clarity in the reconstructed images by optimizing the image gradients, further improving overall image quality. This stage comprehensively evaluates model performance and promotes image restoration, achieving high levels of both visual and structural fidelity.

Hyperparameters were set in order to balance the different levels of concern of the three loss functions. The hyperparameters used in the reconstruction stage for infrared and visible images are: $\alpha_1 = 7$, $\alpha_2 = 4$, $\beta_1 = 6$, and $\beta_2 = 2$. In the first stage, the loss functions used to reconstruct the IR or visible images are L_I^I and L_V^I, respectively, i.e.,

$$L_I^I = L_{MSE}(I, \hat{I}) + \alpha_1 L_{SSIM}(I, \hat{I}) + \alpha_2 L_{Grad}(I, \hat{I}) \tag{3.7}$$

$$L_V^I = L_{MSE}(V, \hat{V}) + \beta_1 L_{SSIM}(V, \hat{V}) + \beta_2 L_{Grad}(V, \hat{V}) \tag{3.8}$$

To further explain the loss function formulas, a detailed description is provided below. In the following three formulas, $\hat{I}^i$ and I^i represent the predicted and true values of the ith pixel, respectively. N denotes the total number of pixels, ∇ represents the gradient to be calculated, and $\|\cdot\|^2$ denotes the L2 norm. These are shown in Eqs (9)–(11), i.e.,

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{I}^i - I^i)^2 \tag{3.9}$$

$$L_{SSIM} = 1 - SSIM(\hat{I}, I) \tag{3.10}$$

$$L_{Grad} = \frac{1}{N}\|\nabla\hat{I}^i - \nabla I^i\|^2 \tag{3.11}$$

The total loss function of the first stage designed by TSTBFuse is L_{total}^I, which is:

$$L_{total}^I = L_I^I + L_V^I \tag{3.12}$$

In the second stage of the total loss function and the first stage of the total loss function, design is similar to the final fusion image obtained in the infrared and visible images, respectively, where MSE loss, SSIM loss, and gradient loss are at the second stage of the total loss function. At the same time, in order to balance the optimization of the loss of the required infrared or visible images, set the hyperparameter as $\eta_1 = 7$, $\eta_2 = 4$, $\gamma_1 = 6$, and $\gamma_2 = 2$, and the total loss function as L_{total}^{II}, i.e.,

$$L_{total}^{II} = L_I^{II} + L_V^{II} \tag{3.13}$$

For the sake of detail, then L1 and L2 are shown in Equations 14 and 15 below, respectively, i.e.,

$$L_I^{II} = L_{MSE}(I, \hat{F}) + \eta_1 L_{SSIM}(I, \hat{F}) + \eta_2 L_{Grad}(I, \hat{F}) \tag{3.14}$$

$$L_V^{II} = L_{MSE}(V, \hat{F}) + \gamma_1 L_{SSIM}(V, \hat{F}) + \gamma_2 L_{Grad}(V, \hat{F}) \tag{3.15}$$

4. Experimentation and analysis

Section 4.1 details the datasets and implementation specifics of TSTBFuse. Section 4.2 presents a comparative analysis, encompassing both qualitative and quantitative evaluations against state-of-the-art algorithms using the TNO, MSRS, and RoadScene datasets. Section 4.3 presents an ablation study on the network architecture and loss function design, demonstrating the performance and robustness of the TSTBFuse model. Finally, Section 4.4 investigates the scalability of the TSTBFuse method.

4.1. Datasets and implementation details

Dataset: 1444 benchmark infrared and visible image pairs in the MSRS [46] dataset were used as the training set for the TSTBFuse experiments, while the other 361 benchmark infrared and visible image pairs in the MSRS dataset were employed as the test set. In addition, 25 pairs and 50 pairs of benchmark datasets in the TNO [47] and RoadScene [48] datasets were selected for testing to verify the generalization performance of the TSTBFuse method, respectively.The MSRS, TNO, and RoadScene datasets are all publicly available datasets, which are available online. Also, the data images used in the TSTBFuse method have been uploaded to GitHub for easy reproduction and use.

Implement details: The training samples were cropped into blocks of 128×128 pixels in steps of 200 to obtain enough training data to train the TSTBFuse method. In this way, 8649 pairs of image blocks were obtained while normalizing their pixel values to [0,1]. During the training and validation process, each pixel value is subtracted by 0.5, which makes the pixel values adjusted to [−0.5,0.5],and improves the learning effect and performance of the model. In the final stage of image reconstruction, the obtained image is added with 0.5, and inverse is normalized to [0,255] to obtain the final fused image. The learning rate is fixed at 0.0001, the batch size is 8, the Adam optimizer is used to optimize the depth model, and the training period (epoch) is 6, where the first and second phases are 3 cycles each. The loss function design hyperparameters are $\alpha_1 = 7$, $\alpha_2 = 4$, $\beta_1 = 6$, $\beta_2 = 2$, $\eta_1 = 7$, $\eta_2 = 4$, $\gamma_1 = 6$, and $\gamma_2 = 2$, which are mainly used to adjust the degree of attention to detail or structure in the final result.

This experiment was conducted under PyTorch 2.0.0 and Python 3.8 with the operating system Ubuntu 20.04. The CUDA version used was 11.8. The hardware configuration consisted of an RTX 4090 graphics card (24 GB), an Intel(R) Xeon(R) Platinum 8352 V processor with 16 virtual CPUs (2.10 GHz), and 90 GB of RAM. This high-performance computing platform provides strong support for the smooth running of the experiment.

4.2. Comparative methods and objective evaluation indicators

In order to validate the importance of the TSTBFuse method, this paper compares seven state-of-the-art algorithms for infrared and visible image fusion, which include the methods DenseFuse [16], deep image decomposition fusion (DIDFuse) [20], nest connection fusion

(NestFuse) [21], and classification saliency fusion (CSFusion) [49] based on the self-encoder structure; the method squeeze-and-decomposition network (SDNet) [50] based on CNNs; the method based on the GAN unified unsupervised fusion (U2Fusion) [51]; and the DATFuse [38] method that combines CNN with Transform. The source code for all seven comparative methods is publicly accessible.

To ensure a comprehensive and objective evaluation, six widely adopted image fusion metrics were employed: information entropy (EN), which quantifies the information richness of the fused images; structural fidelity (SF), which assesses the preservation of structural information, primarily based on gradients or edge information; mutual information (MI), which measures the shared information between two images; structural content distortion (SCD), which evaluates the structural dissimilarity between the reconstructed and reference images; Visual information fidelity (VIF), which compares the visual information transfer capability between images; and the SSIM, which assesses image similarity by comparing luminance, contrast, and structure. SSIM values range from 0 to 1. In this study, the SSIM between the infrared or visible image and the final fused image was calculated separately, and the results were summed for comparison. For all six metrics, higher scores indicate superior fusion performance.

Figure 6 presents four sets of source images from the MSRS validation dataset, along with their corresponding fusion results obtained using TSTBFuse and seven other state-of-the-art methods. To facilitate detailed comparison, local regions of the fused images are magnified to highlight thermal radiation features and finer details. The results indicate that TSTBFuse consistently produces superior fused images compared to the other methods. Specifically, in the magnified regions on the left in the first column, all methods successfully integrated the thermal radiation features from the infrared image. However, the magnified regions on the right reveal that DIDFuse, NestFuse, CSFusion, U2Fusion, and SDNet exhibit limited performance in incorporating details from the visible image. In the magnified regions on the left in the second column, DenseFuse, NestFuse, U2Fusion, SDNet, and DATFuse introduce significant noise during the fusion of infrared thermal radiation features, leading to comparatively blurred results. In the magnified regions on the left of the third and fourth columns, TSTBFuse demonstrates significantly better preservation of IR image features than the other methods. Conversely, in the magnified regions on the right, DIDFuse, NestFuse, and CSFusion produce darker visible image details, and while DATFuse performs better than these, its results remain inferior to TSTBFuse. Therefore, TSTBFuse effectively preserves both the thermal radiation characteristics of the IR image and the details of the visible image within the MSRS test dataset, exhibiting enhanced subjective visual performance compared to the other evaluated methods.

To further provide an objective comparison with the seven other methods on the MSRS test dataset, the same 361 fused infrared and visible image pairs were used to calculate the evaluation metrics for each pair, and the mean values were computed. The results are presented in Table 1. For each metric, higher values indicate better performance, with the best and second-best results highlighted in bold and italics, respectively. As shown in Table 1, TSTBFuse achieved the best performance in EN, SF, SCD, VIF, and SSIM compared to the other methods. While TSTBFuse achieved the second-best result in MI, its performance was significantly better than the other six methods, with DATFuse achieving the best value. Overall, TSTBFuse demonstrated good objective performance, which highlights the effectiveness of its three-branch feature extraction approach.

Figure 7 presents four sets of source images and their corresponding fusion results on the TNO

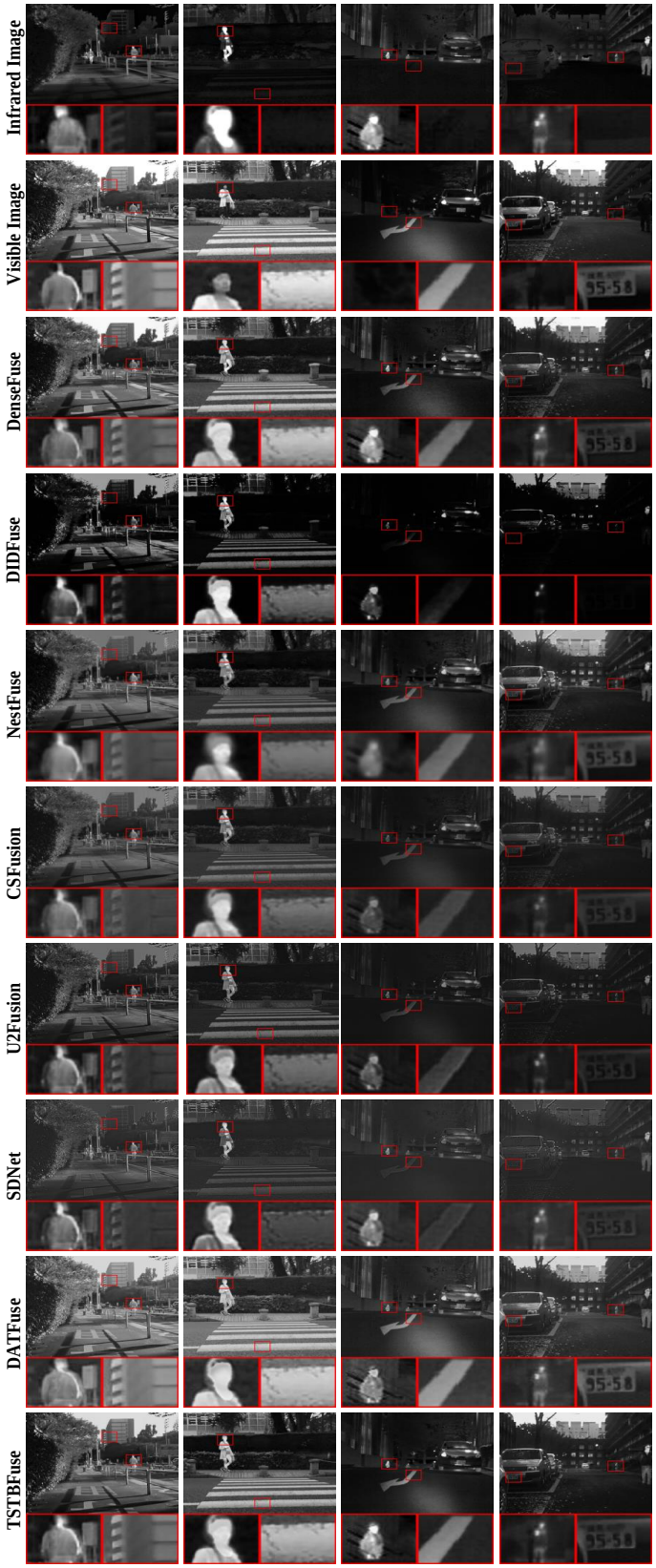

Figure 6. Subjective comparison of TSTBFuse on the MSRS test dataset.

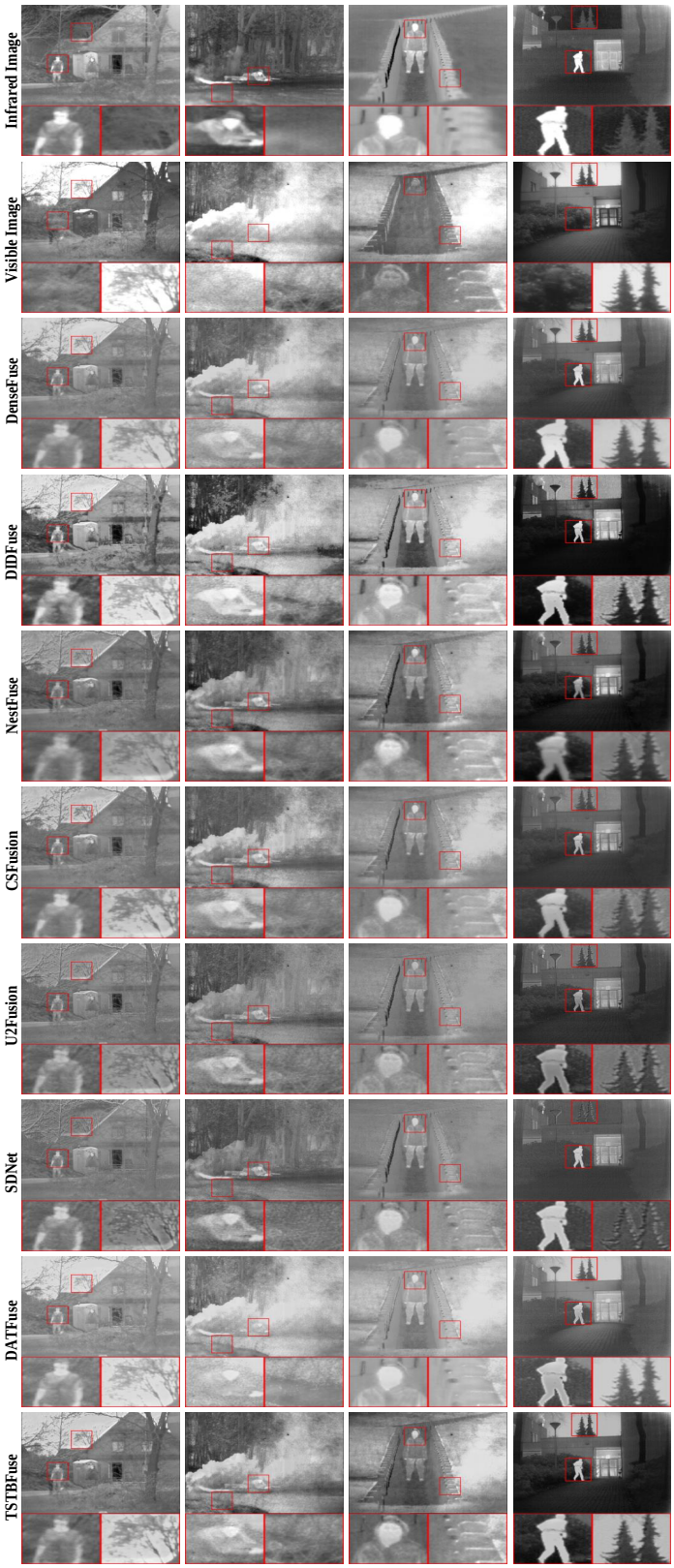

Figure 7. Subjective comparison of TSTBFuse on the TNO test dataset.

Table 1. Presents a quantitative comparison of TSTBFuse's data fusion method with seven state-of-the-art approaches on the MSRS test dataset. For each metric, the best-performing method is highlighted in bold, while the second-best is italicized (to enhance the clarity of data presentation, the DenseFuse, NestFuse, CSFusion, U2Fusion, and DATFuse methods are abbreviated as DenseF, NestF, CSF, U2F, and DATF, respectively. The method names remain consistent across Tables 1–3 and 5).

	EN	SF	MI	SCD	VIF	SSIM
DenseF [16]	6.01	9.27	1.95	1.42	*0.75*	*1.42*
DIDFuse [20]	3.88	9.11	1.30	1.05	0.27	0.60
NestF [21]	*6.21*	6.18	1.71	*1.48*	0.66	1.38
CSF [49]	5.57	5.71	1.65	1.16	0.56	1.39
U2F [51]	5.06	7.26	1.38	1.07	0.49	1.35
SDNet [50]	5.04	7.51	1.18	0.88	0.46	1.18
DATF [38]	6.12	*10.63*	**2.40**	1.41	0.64	1.24
TSTBFuse	**6.31**	**10.78**	*2.03*	**1.55**	**0.76**	**1.46**

dataset, comparing TSTBFuse with seven other state-of-the-art methods. To better illustrate the subjective performance enhancement of TSTBFuse on the TNO dataset, the figure includes local magnifications focusing on target tasks and fine details. In the first column, the magnified details on the right demonstrate that the fused images from DenseFuse, NestFuse, CSFusion, U2Fusion, and SDNet are significantly darker and contain noise or artifacts, resulting in a lower quality compared to TSTBFuse. In the magnified region on the left of the second column, DIDFuse, CSFusion, and DATFuse exhibit excessive haze, leading to poor fusion of target features. Conversely, the magnified regions on the right in the third and fourth columns show that TSTBFuse outperforms the other seven methods in terms of detail and contrast. Furthermore, on the left side, the target task results of DIDFuse and TSTBFuse are also more prominent compared to the other five methods. These results indicate that TSTBFuse achieves superior subjective visual performance compared to the other seven methods.

For an objective comparison using the 25 fused images from the TNO test set, the same six metrics were calculated for each result, and the average value for each metric was computed. The results are shown in Table 2, with the best and second-best values indicated in bold and italics, respectively. As shown in the table, TSTBFuse achieved the best performance in EN, SF, SCD, VIF, and SSIM. Although the TSTBFuse method is slightly lower than DATFuse and DenseFuse in terms of MI metrics, overall it shows good performance in the TNO test data. This is also a side note that the TSTBFuse method has a strong generalization ability and does not need to be trained on the TNO dataset to achieve good results.

Similar to the MSRS and TNO test results, Figure 8 presents a subjective comparison of the source images, TSTBFuse, and the other six methods on the RoadScene dataset. In the right local zoom in the first column, the fusion results of DenseFuse, DIDFuse, CSFusion, U2Fusion, and SDNet are significantly darker and relatively blurred, while the results of the TSTBFuse method are easier to show details. In the local zoom on the right side of the second, third, and fourth columns, it is clear that the TSTBFuse method is more comprehensive in terms of detail presentation. In order to further

Figure 8. Subjective comparison of TSTBFuse on the RoadScene test dataset.

Table 2. Presents a quantitative comparison of TSTBFuse's data fusion method with seven state-of-the-art approaches on the TNO test dataset. For each metric, the best-performing method is highlighted in bold, while the second-best is italicized.

	EN	SF	MI	SCD	VIF	SSIM
DenseF [16]	6.66	8.19	*2.18*	1.52	0.61	1.39
DIDFuse [20]	6.85	9.98	1.69	1.72	0.59	1.17
NestF [21]	*6.89*	5.97	1.5	*1.72*	0.54	1.34
CSF [49]	6.5	6.74	1.47	1.55	0.53	1.40
U2F [51]	6.42	8.85	1.35	1.51	0.54	*1.41*
SDNet [50]	6.36	*10.13*	1.5	1.39	0.53	1.38
DATF [38]	6.58	10.09	**2.36**	1.45	*0.63*	1.33
TSTBFuse	**6.96**	**10.26**	1.8	**1.77**	**0.65**	**1.41**

Table 3. Presents a quantitative comparison of TSTBFuse's data fusion method with seven state-of-the-art approaches on the RoadScene test dataset. For each metric, the best-performing method is highlighted in bold, while the second-best is italicized.

	EN	SF	MI	SCD	VIF	SSIM
DenseF [16]	6.89	9.48	2.08	1.39	0.61	1.02
DIDFuse [20]	7.08	12.08	2.02	1.76	0.61	*1.30*
NestF [21]	*7.09*	7.38	1.89	*1.77*	0.53	1.27
CSF [49]	6.92	8.91	1.86	1.49	0.53	1.15
U2F [51]	6.83	11.13	1.82	1.40	0.56	1.25
SDNet [50]	6.93	*12.10*	2.13	1.24	0.57	1.23
DATF [38]	6.73	11.17	**2.57**	1.29	*0.62*	1.29
TSTBFuse	**7.16**	**12.39**	*2.16*	**1.79**	**0.62**	**1.38**

quantify the performance, the same as the MSRS and TNO test data, the relevant metrics were calculated for 50 fused images of each method and the mean values were taken, and the results are shown in Table 3. Among them, EN, SF, SCD, VIF, and SSIM metrics are better than the other seven methods. Although the TSTBFuse method achieves suboptimal performance in the MI metrics, collectively, it still significantly outperforms the other seven methods in objective comparisons. In addition, the RoadScene test results further indicate that the TSTBFuse method has a strong generalization ability.

Although TSTBFuse performs well on benchmark datasets, we have observed that the model may not achieve optimal performance in cases where visible light images are overexposed or the environment is complex. Under conditions of strong light or low light with noise, the model may overemphasize infrared information, leading to inefficient utilization of visible light textures; in complex scenes with numerous fine edges or overlapping structures, due to the fixed fusion weights in the decoder stage, the fusion output occasionally exhibits excessive smoothing. These limitations reflect the challenges of balancing complementary and shared features in diverse environments. Future research will explore adaptive fusion strategies and local attention mechanisms to address these issues.

Table 4. Quantitative comparison of ablation experiments in the TNO, MSRS, and RoadScene test datasets. For each metric, the best and suboptimal methods are marked with bold and italicized lines, respectively.

Configurations	EN	SF	MI	SCD	VIF	SSIM
(TNO)w/o two-stage training	6.92	10.22	1.80	1.74	0.61	1.38
(TNO)Gradient-free loss function	*6.99*	10.25	1.70	1.75	0.62	*1.40*
(TNO)Balance hyperparameter setting	**7.01**	**10.42**	**2.24**	1.56	**0.72**	1.38
(TNO)TSTBFuse	6.96	*10.26*	*1.80*	**1.77**	*0.65*	**1.41**
(MSRS)w/o two-stage training	6.23	10.45	1.91	1.51	0.70	1.45
(MSRS)Gradient-free loss function	6.30	*11.25*	1.96	*1.51*	0.74	*1.45*
(MSRS)Balance hyperparameter setting	**6.38**	*11.18*	**2.24**	1.36	**0.78**	1.44
(MSRS)TSTBFuse	*6.31*	10.78	*2.03*	**1.55**	0.76	**1.46**
(RoadScene)w/o two-stage training	7.11	*12.18*	2.07	1.70	0.57	1.35
(RoadScene)Gradient-free loss function	7.11	12.16	2.11	*1.76*	0.61	*1.38*
(RoadScene)Hyperparameter calibration	7.10	11.91	**2.39**	1.62	**0.66**	1.30
(RoadScene)TSTBFuse	**7.16**	**12.39**	*2.16*	**1.79**	*0.62*	**1.38**

4.3. Ablation experiments

To further validate the design rationale of the TSTBFuse method, several ablation experiments were conducted. Experiment 1 was designed to evaluate the effectiveness of the two-stage training approach. In this experiment, only the second stage of training was used, while all other settings were kept constant. The experimental results are presented in Table 4: for the TNO dataset, all subjective evaluation metrics were lower compared to those obtained by the full TSTBFuse method. Similarly, across all six evaluation metrics on the MSRS dataset, performance was consistently inferior to that of the TSTBFuse method. For the RoadScene dataset, the objective evaluation metrics also demonstrated a lower performance. These results confirm the effectiveness and rationality of the two-stage training design employed in the TSTBFuse method.

Experiment 2 aimed to evaluate the effectiveness of the proposed loss function design. In this experiment, the gradient loss was removed from both training stages, retaining only the MSE loss and SSIM loss. As shown in Table 4, on the TNO dataset, only the EN metric outperformed the full TSTBFuse method. On the MSRS dataset, the SF metric was the sole indicator exceeding TSTBFuse's performance. Conversely, on the RoadScene dataset, TSTBFuse surpassed the ablation variant across all six evaluation metrics. These results demonstrate the significant contribution of the proposed loss function design to the overall performance enhancement.

Experiment 3 aims to further verify the rationality of the hyperparameters designed in this paper, specifically by balancing the hyperparameters in the loss functions of infrared and visible images. The hyperparameters include include $\alpha_1 = 6, \alpha_2 = 2, \beta_1 = 6, \beta_2 = 2, \eta_1 = 6, \eta_2 = 2, \gamma_1 = 6$, and $\gamma_2 = 2$. The experimental results (see Table 4) show that in the TNO dataset, the EN, SF, MI, and VIF metrics are all higher than the TSTBFuse method; in the MSRS dataset, the EN, SF, MI, and VIF also outperform the TSTBFuse method; whereas in the RoadScene dataset, only the MI and VIF metrics perform better than the TSTBFuse method. Despite the improvement in objective assessment metrics, Experiment 3 fails to retain IR image information well in subjective assessment. In this paper, the fusion results in

the TNO test set are selected from two images for subjective assessment interpretation (see Figure 9), which shows that the final fused image of Experiment III performs poorly in the retention of infrared thermal radiation information. In contrast, the fusion results of the TSTBFuse method achieved the best results in terms of retaining both infrared thermal radiation information and visible image details. This further demonstrates the design rationality and applicability of the TSTBFuse method, which exhibits superior performance in both subjective and objective evaluations.

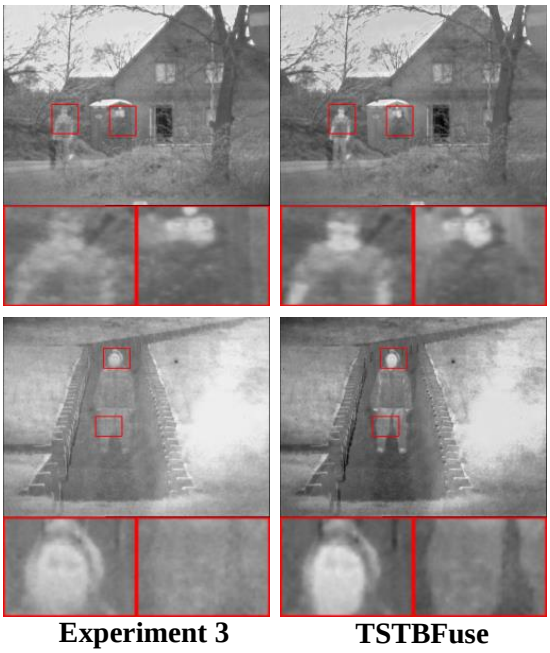

Figure 9. Subjective comparison of results between the ablation experiment (Experiment 3) and TSTBFuse on the TNO test dataset.

Experiment 4 analyzes the sensitivity of key hyperparameters in the loss function. We fixed other parameters while changing one or two parameters, and evaluated performance on the TNO and MSRS datasets using six widely used image fusion metrics (EN, SF, MI, SCD, VIF, and SSIM) (see Tables 5 and 6). Experiments have shown that configurations (7, 4, 6, 2) exhibit strong or optimal performance on multiple metrics on TNO and MSRS datasets, indicating that this setting achieves a good balance between detail preservation, contrast, and structural similarity. When increasing α_1 (the weight of one of the loss terms), such as from 5 to 9, the performance on EN slightly improves (TNO: 6.72 $\rightarrow$ 7.02), but at the cost of sacrificing VIF and SSIM. Similarly, excessive increase in α_2 may lead to a decrease in MI. These results support the importance of carefully adjusting loss weights and validate our final configuration based on equilibrium trade-offs.

4.4. Extensibility of the methodology

To further assess the generalization ability of the TSTBFuse method in different domains, this study selected the multimodal medical image dataset MRI-CT for validation. The pretrained model

Table 5. Results of ablation experiments with different parameter weights in the test set of TNO. Bold indicates the best value.

Configurations: $(\alpha_1, \alpha_2, \beta_1, \beta_2)$	EN	SF	MI	SCD	VIF	SSIM
(5, 4, 6, 2)	6.72	9.62	1.73	1.25	0.44	1.18
(9, 4, 6, 2)	**7.02**	10.03	1.66	0.17	0.33	0.98
(7, 2, 6, 2)	6.97	10.22	1.82	0.56	0.44	1.09
(7, 6, 6, 2)	6.9	10.21	**2.49**	0.39	0.65	1.21
Ours (7, 4, 6, 2)	6.69	**10.26**	1.8	**1.77**	**0.65**	**1.41**

Table 6. Results of ablation experiments with different parameter weights in the test set of MSRS. Bold indicates the best value.

Configurations: $(\alpha_1, \alpha_2, \beta_1, \beta_2)$	EN	SF	MI	SCD	VIF	SSIM
(5, 4, 6, 2)	6.36	8.81	1.54	0.7	0.38	0.82
(9, 4, 6, 2)	**6.83**	7.21	1.86	0.39	0.55	0.66
(7, 2, 6, 2)	5.95	7.46	1.57	0.32	0.41	0.72
(7, 6, 6, 2)	6.49	9.68	1.84	0.86	0.62	0.96
Ours (7, 4, 6, 2)	6.31	**10.78**	**2.03**	**1.55**	**0.76**	**1.46**

parameters were directly applied without any fine-tuning or retraining, and testing was performed on 21 image pairs. For subjective evaluation, TSTBFuse was compared with DenseFuse, DIDFuse, NestFuse, CSFusion, U2Fusion, SDNet, and DATFuse, with the results presented in Figure 10. It is evident that in the first and fourth columns, NestFuse, CSFusion, U2Fusion, and SDNet exhibit poor detail in the left magnified regions, whereas TSTBFuse demonstrates richer detail preservation. In the second and third columns, the contrast features extracted by DIDFuse, NestFuse, and SDNet appear significantly darker and less distinct compared to those of TSTBFuse. For objective evaluation, six metrics—EN, SF, MI, SCD, VIF, and SSIM—were used to compare TSTBFuse against the other seven methods, with results summarized in Table 7. The data shows that TSTBFuse has significant improvement in three indicators, including SCD, VIF, and SSIM. Although its performance is slightly inferior to DATFuse in EN, SF, and MI, overall, TSTBFuse demonstrates superior effectiveness in the direct evaluation of this dataset.

To further validate the generalization performance of the TSTBFuse method, this study directly tested fusion on 50 pairs of RGB and infrared images from the RoadScene dataset. Since TSTBFuse is designed to accept single-channel inputs, the RGB images were converted to YCbCr (luminance-chrominance) color space during testing. Specifically, the Y channel was used as the visible image input to the model, while the Cb and Cr channels were retained. After processing by the model, a single-channel fused image was produced, which was then combined with the original Cb and Cr channels, converted back to YCbCr format, and finally transformed into an RGB image for saving. The results show that the TSTBFuse method can directly fuse RGB color images with IR images without fine-tuning and retraining. Figure 11 illustrates the fusion results of 10 representative images from the RoadScene test dataset.

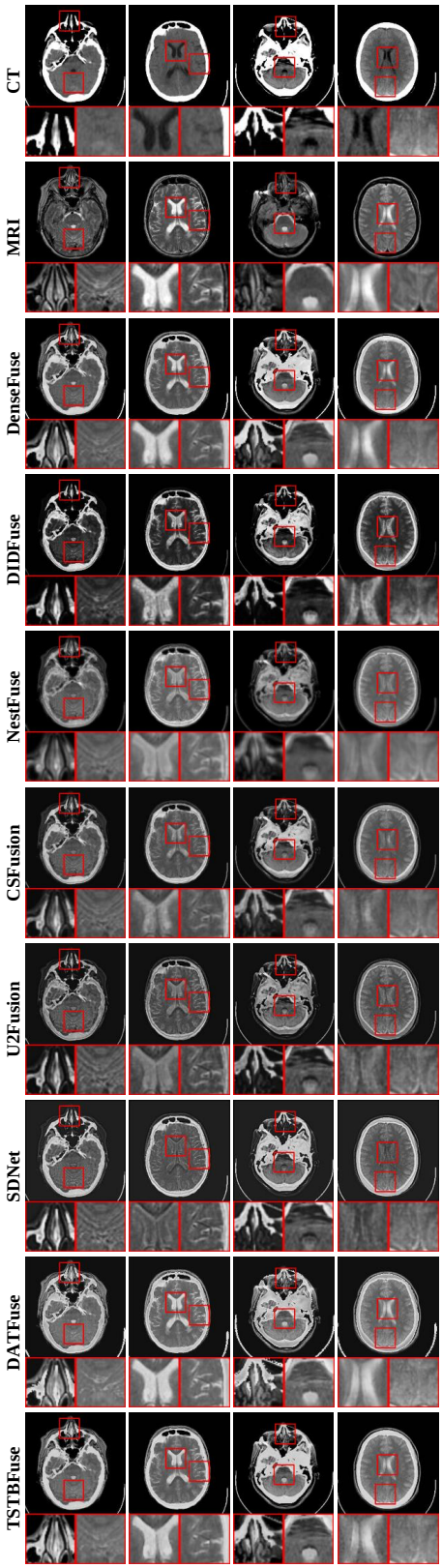

Figure 10. Subjective comparison of TSTBFuse on the MRI-CT test dataset.

Figure 11. Subjective comparison of TSTBFuse on the RoadScene test dataset.

Table 7. Quantitative comparison of the scalability of the TSTBFuse method with seven other methods in the MRI-CT test dataset. For each metric, the best and suboptimal methods are marked with bold and italicized lines, respectively.

	EN	SF	MI	SCD	VIF	SSIM
DenseF [16]	4.71	*30.05*	*2.36*	*1.21*	*0.44*	1.4
DIDFuse [20]	4.40	25.15	2.24	0.69	0.39	1.28
NestF [21]	4.82	10.86	2.12	1.08	0.37	*1.41*
CSF [49]	4.75	16.83	2.23	0.8	0.41	0.6
U2F [51]	4.78	20.44	2.09	0.52	0.36	0.57
SDNet [50]	4.84	23.84	2.12	0.43	0.38	0.55
DATF [38]	**5.06**	**33.98**	**2.39**	1.08	0.42	0.55
TSTBFuse	*4.85*	26.16	2.23	**1.42**	**0.47**	**1.43**

5. Conclusions

In this paper, we propose a TSTBFuse for infrared and visible images based on a self-encoder structure, which is an end-to-end model where the fusion layer does not need to manually design the fusion rules. In a specifically designed three-branch feature extraction encoder network, TSTBFuse is able to efficiently extract infrared features, visible features, and shared features. Furthermore, a combined loss function—comprising MSE loss, SSIM loss, and gradient loss—is employed to effectively preserve thermal radiation information from infrared images and detail information from visible images in a supervised manner. Extensive subjective and objective evaluations on three publicly available datasets—TNO, MSRS, and RoadScene—demonstrate that TSTBFuse outperforms seven state-of-the-art fusion methods across multiple metrics. In addition, the method was successfully extended to the multimodal medical image dataset MRI-CT for testing, showing good generalization capabilities and the ability to generate RGB image results directly during the testing phase. Although the current model performs well in terms of fusion effectiveness, the three-branch structure and Transformer decoder result in high computational overhead, leading to relatively long inference times, which means it does not yet fully meet the real-time requirements of scenarios such as intelligent driving. This trade-off between fusion performance and runtime efficiency will be a key focus of future optimization efforts. Future work will focus on developing a real-time fusion method for infrared and RGB color images, targeting the challenges of dynamic scene fusion and enabling instantaneous fusion for the same scene. Furthermore, we plan to integrate the model into downstream tasks such as semantic segmentation, object detection, and autonomous driving perception to further validate its practical application value and contribution to real-world visual systems.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This study was financially supported by the following projects: Science and Technology Research Project of Henan Province (No. 242102211110, No. 242102210217). The data in this study are available on request to the second author.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. Y. Kuai, D. Li, Z. Gao, M. Yuan, D. Zhang, Visible-infrared dual-sensor tracking based on transformer via progressive feature enhancement and fusion, *IEEE Sens. J.*, **24** (2024), 14519–14528. https://doi.org/10.1109/JSEN.2024.3372991
2. J. Wang, L. Chu, C. Guo, Y. Zhang, Z. Cao, Target track enhancement based on asynchronous radar and camera fusion in intelligent driving system, *IEEE Sens. J.*, **24** (2023), 3131–3143. https://doi.org/10.1109/JSEN.2023.3339328
3. H. Qin, X. Zhang, R. Gong, Y. Ding, Y. Xu, X. Liu, Distribution-sensitive information retention for accurate binary neural network, *Int. J. Comput. Vision*, **131** (2023), 26–47. https://doi.org/10.1007/s11263-022-01687-5
4. Q. Zhang, T. Xiao, N. Huang, D. Zhang, J. Han, Revisiting feature fusion for RGB-T salient object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **31** (2020), 1804–1818. https://doi.org/10.1109/TCSVT.2020.3014663
5. X. Li, C. Luo, J. Dezert, Y. Tan, Generic object recognition based on feature fusion in robot perception, *Int. J. Rob. Autom.*, **31** (2016), 1–7. https://dx.doi.org/10.2316/Journal.206.2016.5.206-4706
6. N. Aldahoul, H. A. Karim, M. A. Momo, F. I. F. Escobara, M. J. T. Tan, Space object recognition with stacking of CoAtNets using fusion of RGB and depth images, *IEEE Access*, **11** (2023), 5089–5109. https://doi.org/10.1109/ACCESS.2023.3235965
7. C. Thomas, T. Ranchin, L. Wald, J. Chanussot, Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics, *IEEE Trans. Geosci. Remote Sens.*, **46** (2008), 1301–1312. https://doi.org/10.1109/TGRS.2007.912448
8. Y. Liu, X. Chen, R. K. Ward, Z. J. Wang, Image fusion with convolutional sparse representation, *IEEE Signal Process. Lett.*, **23** (2016), 1882–1886. https://doi.org/10.1109/LSP.2016.2618776
9. A. Dogra, B. Goyal, S. Agrawal, From multi-scale decomposition to non-multi-scale decomposition methods: A comprehensive survey of image fusion techniques and its applications, *IEEE Access*, **5** (2017), 16040–16067. https://doi.org/10.1109/ACCESS.2017.2735865
10. E. Karami, S. Prasad, M. Shehata, Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images, preprint, arXiv:1710.02726. https://doi.org/10.48550/arXiv.1710.02726

11. J. A. Aghamaleki, A. Ghorbani, Image fusion using dual tree discrete wavelet transform and weights optimization, *Visual Comput.*, **39** (2023), 1181–1191. https://doi.org/10.1007/s00371-021-02396-9
12. Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, *Inf. Fusion*, **24** (2015),147–164. https://doi.org/10.1016/j.inffus.2014.09.004
13. H. Ghassemian, A review of remote sensing image fusion methods, *Inf. Fusion*, **32** (2016), 75–89. https://doi.org/10.1016/j.inffus.2016.03.003
14. J. Fu, W. Li, J. Du, B. Xiao, Multimodal medical image fusion via laplacian pyramid and convolutional neural network reconstruction with local gradient energy strategy, *Comput. Biol. Med.*, **126** (2020), 104048. https://doi.org/10.1016/j.compbiomed.2020.104048
15. J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, *Inf. Fusion*, **45** (2019), 153–178. https://doi.org/10.1016/j.inffus.2018.02.004
16. H. Li, X. J. Wu, DenseFuse: A fusion approach to infrared and visible images, *IEEE Trans. Image Process.*, **28** (2018), 2614–2623. https://doi.org/10.1109/TIP.2018.2887342
17. H. Li, X. J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in *2018 24th International Conference on Pattern Recognition (ICPR)*, (2018), 2705–2710. https://doi.org/10.1109/ICPR.2018.8546006
18. J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, *Inf. Fusion*, **48** (2019), 11–26. https://doi.org/10.1016/j.inffus.2018.09.004
19. J. Ma, H. Zhang, Z. Shao, P. Liang, H. Xu, GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion, *IEEE Trans. Instrum. Meas.*, **70** (2020), 1–14. https://doi.org/10.1109/TIM.2020.3038013
20. Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, J. Zhang, DIDFuse: Deep image decomposition for infrared and visible image fusion, preprint, arXiv:2003.09210. https://doi.org/10.48550/arXiv.2003.09210
21. H. Li, X. J. Wu, T. Durrani, NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, *IEEE Trans. Instrum. Meas.*, **69** (2020), 9645–9656. https://doi.org/10.1109/TIM.2020.3005230
22. N. Aishwarya, C. B. Thangammal, Visible and infrared image fusion using DTCWT and adaptive combined clustered dictionary, *Infrared Phys. Technol.*, **93** (2018), 300–309. https://doi.org/10.1016/j.infrared.2018.08.013
23. J. Li, J. Liu, S. Zhou, Q. Zhang, N. K. Kasabov, Learning a coordinated network for detail-refinement multiexposure image fusion, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2022), 713–727. https://doi.org/10.1109/TCSVT.2022.3202692
24. F. P. An, X. M. Ma, L. Bai, Image fusion algorithm based on unsupervised deep learning-optimized sparse representation, *Biomed. Signal Process. Control*, **71** (2022), 103140. https://doi.org/10.1016/j.bspc.2021.103140
25. H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, *Inf. Fusion*, **76** (2021), 323–336. https://doi.org/10.1016/j.inffus.2021.06.008

26. R. C. Gonzalez, *Digital Image Processing*, Pearson Education India, 2009. https://doi.org/10.1117/1.3115362

27. G. M. Foody, Sharpening fuzzy classification output to refine the representation of sub-pixel land cover distribution, *Int. J. Remote Sens.*, **19** (1998), 2593–2599. https://doi.org/10.1080/014311698214659

28. P. J. Burt, E. H. Adelson, The Laplacian pyramid as a compact image code, *Read. Comput. Vision*, **1987** (1987), 671–679. https://doi.org/10.1016/B978-0-08-051581-6.50065-9

29. L. Li, S. Jiang, Q. Huang, Learning image Vicept description via mixed-norm regularization for large scale semantic image search, *CVPR 2011*, **2011** (2011), 825–832. https://doi.org/10.1109/CVPR.2011.5995570

30. M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, et al., A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, *Comput. Biol. Med.*, **144** (2022), 105253. https://doi.org/10.1016/j.compbiomed.2022.105253

31. S. Kavitha, K. K. Thyagharajan, Efficient DWT-based fusion techniques using genetic algorithm for optimal parameter estimation, *Soft Comput.*, **21** (2017), 3307–3316. https://doi.org/10.1007/s00500-015-2009-6

32. M. Yin, W. Liu, X. Zhao, Y. Yin, Y. Guo, A novel image fusion algorithm based on nonsubsampled shearlet transform, *Optik*, **125** (2014), 2274–2282. https://doi.org/10.1016/j.ijleo.2013.10.064

33. Y. Li, H. Zhao, Z. Hu, Q. Wang, Y. Chen, IVFuseNet: Fusion of infrared and visible light images for depth prediction, *Inf. Fusion*, **58** (2020), 1–12. https://doi.org/10.1016/j.inffus.2019.12.014

34. Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, IFCNN: A general image fusion framework based on convolutional neural network, *Inf. Fusion*, **54** (2020), 99–118. https://doi.org/10.1016/j.inffus.2019.07.011

35. J. Yue, L. Fang, S. Xia, Y. Deng, J. Ma, Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models, *IEEE Trans. Image Process.*, **32** (2023), 5705–5720. https://doi.org/10.1109/TIP.2023.3322046

36. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 4700–4708. https://doi.org/10.1109/CVPR.2017.243

37. H. Xu, M. Gong, X. Tian, J. Huang, J. Ma, CUFD: An encoder-decoder network for visible and infrared image fusion based on common and unique feature decomposition, *Comput. Vision Image Understanding*, **218** (2022), 103407. https://doi.org/10.1016/j.cviu.2022.103407

38. W. Tang, F. He, Y. Liu, Y. Duan, T. Si, DATFuse: Infrared and visible image fusion via dual attention transformer, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 3159–3172. https://doi.org/10.1109/TCSVT.2023.3234340

39. M. Xia, C. Lin, B. Xu, Q. Li, H. Fang, Z. Huang, DSAFusion: Detail-semantic-aware network for infrared and low-light visible image fusion, *Infrared Phys. Technol.*, **147** (2025), 105804. https://doi.org/10.1016/j.infrared.2025.105804

40. Z. Huang, C. Lin, B. Xu, M. Xia, Q. Li, Y. Li and N. Sang, T^{2}EA: Target-aware Taylor expansion approximation network for infrared and visible image fusion, *IEEE Trans. Circuits Syst. Video Technol.*, **35** (2025), Forthcoming.
41. Z. Huang, C. Lin, B. Xu, M. Xia, Q. Li, N. Sang, MSCS: Multi-stage feature learning with channel-spatial attention mechanism for infrared and visible image fusion, *Infrared Phys. Technol.*, **142** (2024), 105514. https://doi.org/10.1016/j.infrared.2024.105514
42. R. Hou, VIF-Net: An unsupervised framework for infrared and visible image fusion, *IEEE Trans. Comput. Imaging*, **6** (2020), 640–651. https://doi.org/10.1109/TCI.2020.2965304
43. Y. Liu, Y. Zang, D. Zhou, J. Cao, R. Nie, R. Hou, et al., An improved hybrid network with a Transformer module for medical image fusion, *IEEE J. Biomed. Health Inf.*, **27** (2023), 3489–3500. https://doi.org/10.1109/JBHI.2023.3264819
44. Z. Ding, H. Li, D. Zhou, Y. Liu, R. Hou, A robust infrared and visible image fusion framework via multi-receptive-field attention and color visual perception, *Appl. Intell.*, **53** (2023), 8114–8132. https://doi.org/10.1007/s10489-022-03952-z
45. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594
46. L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, PIAFusion: A progressive infrared and visible image fusion network based on illumination aware, *Inf. Fusion*, **83** (2022), 79–92. https://doi.org/10.1016/j.inffus.2022.03.007
47. A. Toet, M. A. Hogervorst, Progress in color night vision, *Optical Eng.*, **51** (2012), 010901. https://doi.org/10.1117/1.OE.51.1.010901
48. H. Xu, J. Ma, Z. Le, J. Jiang, X. Guo, Fusiondn: A unified densely connected network for image fusion, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 12484–12491. https://doi.org/10.1609/aaai.v34i07.6936
49. H. Xu, H. Zhang, J. Ma, Classification saliency-based rule for visible and infrared image fusion, *IEEE Trans. Comput. Imaging*, **7** (2021), 824–836. https://doi.org/10.1109/TCI.2021.3100986
50. H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, *Int. J. Comput. Vision*, **129** (2021), 2761–2785. https://doi.org/10.1007/s11263-021-01501-8
51. H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2020), 502–518. https://doi.org/10.1109/TPAMI.2020.3012548