# YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer

Wei Tang, Fazhi He, *Member, IEEE*, and Yu Liu, *Member, IEEE*

***Abstract*—Infrared and visible image fusion is aims to generate a composite image that can simultaneously describe the salient target in the infrared image and texture details in the visible image of the same scene. Since deep learning (DL) exhibits great feature extraction ability in computer vision tasks, it has also been widely employed in handling infrared and visible image fusion issue. However, the existing DL-based methods generally extract complementary information from source images through convolutional operations, which results in limited preservation of global features. To this end, we propose a novel infrared and visible image fusion method, i.e., the Y-shape dynamic Transformer (YDTR). Specifically, a dynamic Transformer module (DTRM) is designed to acquire not only the local features but also the significant context information. Furthermore, the proposed network is devised in a Y-shape to comprehensively maintain the thermal radiation information from the infrared image and scene details from the visible image. Considering the specific information provided by the source images, we design a loss function that consists of two terms to improve fusion quality: a structural similarity (SSIM) term and a spatial frequency (SF) term. Extensive experiments on mainstream datasets illustrate that the proposed method outperforms both classical and state-of-the-art approaches in both qualitative and quantitative assessments. We further extend the YDTR to address other infrared and RGB-visible images and multi-focus images without fine-tuning, and the satisfactory fusion results demonstrate that the proposed method has good generalization capability.**

***Index Terms*—Dynamic transformer, image fusion, infrared image, Y-shape network.**

## I. Introduction

DUE to the difference in imaging sensors, infrared images and visible images reveal different characteristics [1]. The infrared image can detect targets well under all weather conditions, since the infrared sensor identifies objects by thermal radiation. However, infrared images provide limited texture details of scenes due to low spatial resolution [2]. By comparison, the visible image reports high spatial resolution and is more consist with human eyes, while it will be blurry at night or under bad weather conditions [3]. The fusion of infrared and visible images can generate a single image that provides prominent targets with a clear background and facilitates subsequent vision tasks, such as military applications [4], depth prediction [5], and face recognition [6], [7].

During the past decades, numerous infrared and visible image fusion methods hava been proposed, which can be roughly divided into two categories: conventional methods [1], [8]–[17] and deep learning (DL)-based methods [2], [18]–[24]. The conventional methods primarily involve multi-scale transform (MST)-based methods [8], [9], [11], sparse representation (SR)-based methods [12], [15], total variation (TV)-based methods [10], [14], hybrid methods [13], [16], and other methods [1], [17]. Although conventional methods have achieved good fusion results, these methods are becoming increasingly complicated and time-consuming for pursuing better fusion performance. In addition, the conventional methods need manually designing decomposition approaches and fusion strategies, by doing so, significant complementary information of the source images can hardly be comprehensively preserved. Therefore, DL-based methods have emerged as a hot topic in addressing infrared and visible image fusion task during the past few years on account of the powerful feature representation capability. DL-based methods mainly include convolutional neural network (CNN)-based methods [18], [20], [22], [23] and generative adversarial network (GAN)-based methods [2], [19], [21], [24]. However, the existing DL-based methods generally employ convolutional operations for feature extraction, which can well capture local features while achieving limited performance in maintaining the global context information from the source images. Moreover, most of these DL-based algorithms first concatenate the source images, and subsequently feed them into a single-path network without separate feature extraction branches for specific input images, which causes the loss of some unique information from the source images.

To address these limitations, this paper proposes a novel end-to-end infrared and visible image fusion method via a Y-shape dynamic Transformer called YDTR. YDTR consists of two Y branches and a main path. The two Y branches are designed to separately extract thermal radiation information from the infrared image and texture details from the visible image. Each branch involves an encoder and a dynamic Transformer module (DTRM), where the encoder is devised to capture shallow

Manuscript received 4 March 2022; revised 6 June 2022, 12 July 2022, and 14 July 2022; accepted 16 July 2022. Date of publication 20 July 2022; date of current version 30 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072348 and 62176081, in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, and in part by the National Key R&D Program of China under Grant 2018AAA0101104. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Erkut Erdem. (*Corresponding author: Fazhi He.*)

Wei Tang and Fazhi He are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: weitang2021@whu.edu.cn; fzhe@whu.edu.cn).

Yu Liu is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: yuliu@hfut.edu.cn).

Our code is available at https://github.com/tthinking/YDTR.

Digital Object Identifier 10.1109/TMM.2022.3192661

1520-9210 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

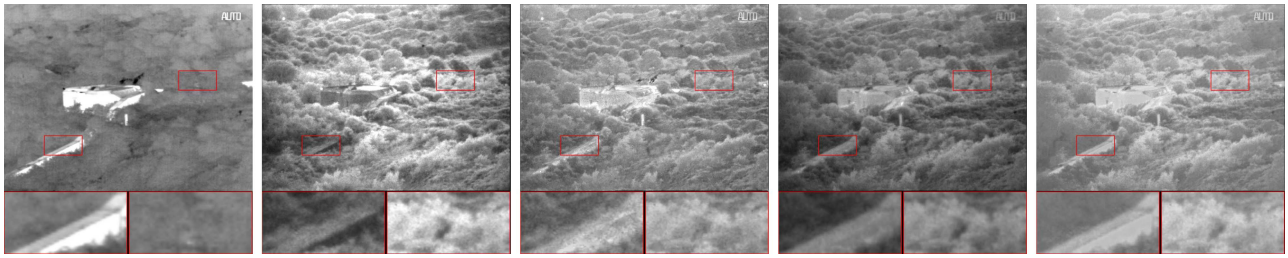


Fig. 1. Schematic illustration of infrared and visible image fusion. From left to right: infrared image, visible image, fusion results of RP [8], RFN-Nest [23], and the proposed YDTR.

features from the source images while the DTRM is utilized to model long-range complementary relationships. The main path is composed of a DTRM for feature integration and a decoder for dimension reduction. Since the proposed method is trained in an unsupervised manner, to further improve the fusion quality, we introduce a new loss function for network training. Specifically, a structural similarity (SSIM) term and a spatial frequency (SF) term are presented for loss function construction to preserve more significant information from the source images.

To illustrate the effectiveness of the proposed method, Fig. 1 reports a pair of infrared and visible images and the corresponding fusion results of a representative conventional method RP [8], a recently proposed DL-based method RFN-Nest [23] and the proposed YDTR. In each image, two local regions are enlarged as close-ups for better comparison. The infrared image highlights the objects, e.g., the building and road, however, the environment around the targets is unclear. The visible image provides more detailed information, while the objects are not salient. Both RP-based and RFN-Nest-based methods achieve good fusion performance in terms of maintaining the complementary information from the source images. Nevertheless, compared to the proposed method, these two competitors tend to be blurry to some degree. More specifically, the road in the YDTR is distinct, while it is difficult to identify whether there is a road in the RP-based and RFN-Nest-based methods (see the first close-ups in Fig. 1). In addition, the RP-based and RFN-Nest-based methods lose some texture details, which results in unnatural contrast to some degree (see the second close-ups in Fig. 1). Overall, the proposed YDTR exhibits satisfactory performance in simultaneously exploring the thermal radiation information from the infrared image and scene details from the visible image.

The main contributions of this paper can be summarized as the following five points.

- Unlike most existing DL-based image fusion methods that extract features through convolutional operations, we propose a dynamic Transformer for long-term relationship construction so that not only the local useful information but also the global complementary information can be fully integrated.
- To fully exploit significant information from the source images, the proposed network is designed in a Y-shape fashion. The two Y shaped branches are designed to separately preserve the thermal radiation information of the infrared image and texture details of the visible image. The main path is devised to adequately merge the extracted features.
- Since there is lack of ground-truth for infrared and visible image fusion, the loss function plays a crucial role in affecting the fusion performance. To this end, we design a new loss function that consists of an SSIM term and an SF term to train the proposed network in an unsupervised manner so that more salient features and detailed information can be maintained in the fusion result.
- Extensive experiments on two mainstream datasets demonstrate that the proposed method outperforms the other nine representative and state-of-the-art methods in both quantitative and qualitative assessments.
- We also extend the proposed method to address the infrared and RGB-visible image fusion and multi-focus image fusion tasks without fine-tuning. The satisfactory generalization results illustrate that our YDTR has good generalization ability.

The remainder of this paper is arranged as follows. Section II introduces some related works and the motivations of this work. The proposed method is described in Section III. Experimental details are provided in Section IV. Finally, this paper is concluded in Section V.

## II. Related Works and Motivations

In this section, the related works on infrared and visible image fusion are first reviewed in Section II-A, followed by a brief introduction to the Transformer in Section II-B. Finally, the motivations of this work are illustrated in Section II-C.

### A. Infrared and Visible Image Fusion

Many infrared and visible image fusion methods have been proposed in the past decades [3], among which MST-based methods are the most representative [13]. Generally, MST-based methods consist of three main steps [34]–[36]: (i) The source images are decomposed to obtain multi-scale sub-bands. (ii) The coefficients are merged by specific fusion strategies. (iii) The fused coefficients are reconstructed to generate the fusion result. Obviously, decomposition approaches and fusion rules are two crucial parts that determine fusion performance. However, to improve the fusion quality, these two key steps tend to be increasingly complex, which drains time and resources.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

To address these problems, benefitting from the powerful feature extraction capability, DL-based infrared and visible image fusion methods have become a popular theme in the last few years [2], [18]–[20], [23], [24]. Liu *et al.* [25] first applied a CNN to infrared and visible image fusion field, where a Siamese convolutional network was utilized to achieve activity level measurement and weight assignment as a whole. Zhao *et al.* [26] designed an infrared and visible image fusion method (AUIF) by combining model-based prior information and traditional two-scale decomposition approaches. Li *et al.* [18] presented an infrared and visible image fusion approach based on CNN and dense connection, called DenseFuse. In 2019, Ma *et al.* [19] first adopted GAN to address the image fusion task, named FusionGAN. In their method, an adversarial game is established between a generator and a discriminator to obtain the fusion result. Li *et al.* [2] developed FusionGAN by integrating a multi-scale attention mechanism into both generator and discriminator to perceive more discriminative regions, called AttentionFGAN. Ma *et al.* [24] further improved FusionGAN and presented a generative adversarial network with multiclassification constraints (GANMcC) for infrared and visible image fusion, which transforms image fusion into a multi-distribution simultaneous estimation problem. In [27], Li *et al.* also employed a GAN-based model to address infrared and visible image fusion task. They introduced a multigrained attention module into an encoder-decoder network to fuse infrared and visible images (MgAN-Fuse). Xu *et al.* [20] presented a unified image fusion network for several image fusion tasks, including infrared and visible images, by employing continuous learning to train a single model, which avoided the problem of catastrophic forgetting, storage, and computation. In 2021, Li *et al.* [23] put forward a residual fusion network for infrared and visible image fusion with a two-stage training strategy. Liu *et al.* [28] presented a deep network for infrared and visible image fusion, where a feature learning module with a fusion learning mechanism was cascaded. Xu *et al.* [29] introduced a pixel-wise classification saliency-based fusion rule for infrared and visible image fusion. In their method, a classifier is employed to classify two types of source images, and the importance of each pixel is quantified as its contribution to the classification result.

Although DL-based methods have achieved better fusion performance, they still suffer from two key drawbacks. (i) The features are captured through convolutional operations, which cannot well model long-range context information, and results in the loss of some significant global features. (ii) Single network or two parallel networks are adopted for feature extraction without careful consideration of specific information that existed in the source images.

### B. Transformer

In 2017, Vaswani *et al.* [37] first presented the concept of the Transformer, which conquers the inherent problem in CNN, i.e., restricted receptive field, by employing multi-head self-attention. Since then, the Transformer has been widely used in the field of natural language processing (NLP) [38]–[40]. With the successful application in NLP, Dosovitskiy *et al.* [41] employed the Transformer for image classification, which is called Vision Transformer (ViT). Benefiting from its promising global context feature exploration capability, many Transformer-based methods have been proposed to handle computer vision tasks. Zheng *et al.* [42] deployed a pure transformer for semantic segmentation, where the semantic segmentation was treated as a sequence-to-sequence prediction task. Wang *et al.* [43] put forward a Pyramid vision Transformer (PVT) for various dense prediction tasks, which inherits the merits of both CNN and Transformer. Chen *et al.* [44] proposed a TransUNet for medical image segmentation by meriting both Transformers and U-Net. Wang *et al.* [45] introduced a video instance segmentation framework based on Transformers. In their method, video instance segmentation is viewed as a direct end-to-end parallel sequence prediction problem.

Most recently, several Transformer-based image fusion methods have been proposed. In [30], an image fusion Transformer is presented. Their proposed method follows a two-stage training approach by adopting a transformer-based multi-scale fusion strategy. Zhao *et al.* [31] presented a DenseNet-Transformer for infrared and visible image fusion, where DenseNet was utilized as the encoder and a dual-Transformer was employed as a fusion strategy. Subsequently, Fu *et al.* [32] put forward a patch Pyramid Transformer (PPT) for image fusion. In their method, a patch Transformer was devised to transform the image into a sequence of patches, and a Pyramid Transformer is designed for feature extraction. In [33], Rao *et al.* presented an infrared and visible image fusion approach by combining Transformer and adversarial learning, where a Transformer-based generator and two discriminators were devised.

Motivated by the pleasing long-range dependency modeling capability of Transformer, in this work, we propose a new infrared and visible image fusion method by combining the advantages of both CNN and Transformer. In this way, the complementary local and global features can be comprehensively preserved from the source images.

### C. Motivations

Infrared and visible image fusion aims to obtain a composite image that can simultaneously maintain the thermal radiation information from the infrared image and texture details from the visible image to provide salient objects with clear scenes. With this consideration, we propose an end-to-end and unsupervised Y-shape dynamic Transformer fusion model. Specifically, to avoid manual design like traditional methods, the proposed method is devised in an end-to-end manner so that fusion results can be generated automatically. Besides, since convolutional operation has limited global feature extraction capability, we integrate the advantages of both CNN and Transformer, and design a dynamic Transformer to achieve local feature preservation and long-range relationship construction. In addition, to fully extract the complementary information from the source images, two branches are presented in a Y-shape to separately exploit specific features. Furthermore, as there is lack of ground-truth for infrared and visible image fusion, the loss function plays an important role in influencing the fusion results. To this end,

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON WITH STATE-OF-THE-ART IMAGE FUSION ALGORITHMS

| Methods | End-to-End | Convolutional Operation | Transformer | Y-shape | SSIM Loss | SF Loss | Unsupervised | Generalization Ability |
|---|---|---|---|---|---|---|---|---|
| CNN [25] | × | ✓ | × | × | × | × | × | × |
| AUIF [26] | × | ✓ | × | × | ✓ | × | ✓ | × |
| DenseFuse [18] | × | ✓ | × | × | ✓ | × | ✓ | × |
| FusionGAN [19] | ✓ | ✓ | × | × | × | × | ✓ | × |
| AttentionFGAN [2] | ✓ | ✓ | × | ✓ | × | × | ✓ | × |
| GANMcC [24] | ✓ | ✓ | × | × | × | × | ✓ | ✓ |
| MgAN-Fuse [27] | ✓ | ✓ | × | × | × | × | ✓ | × |
| U2Fusion [20] | ✓ | ✓ | × | × | ✓ | × | ✓ | × |
| RFN-Nest [23] | ✓ | ✓ | × | × | ✓ | × | × | ✓ |
| MFE-EAG [28] | ✓ | ✓ | × | × | ✓ | × | × | ✓ |
| CSF [29] | × | ✓ | × | × | ✓ | × | ✓ | × |
| IFT [30] | × | ✓ | ✓ | × | ✓ | × | ✓ | × |
| DNDT [31] | ✓ | ✓ | ✓ | × | ✓ | × | ✓ | × |
| PPT Fusion [32] | × | × | ✓ | × | × | × | × | × |
| TGFuse [33] | ✓ | ✓ | ✓ | × | ✓ | × | ✓ | × |
| YDTR | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

an SSIM loss and an SF loss are presented and combined as the objective function to train the proposed fusion model in an unsupervised fashion.

Table I comprehensively compare the proposed method with some state-of-the-art image fusion algorithms from eight different aspects.

- *End-to-End*. One superiority of DL-based image fusion methods over traditional methods is that DL-based fusion models can be designed in an end-to-end manner to directly obtain fused images without elaborately devised fusion strategies.
- *Convolutional Operation*. The convolutional operation can improve the fusion performance on account of its powerful feature representation ability. However, several recently presented Transformer-based fusion models only utilize the pure Transformer, which may result in inadequate preservation of local significant features.
- *Transformer*. Since the CNN has limited receptive fields, it cannot well model long-range dependencies, and Transformer embedded model can alleviate this drawback and construct global relationships.
- *Y-shape*. A single network can produce good fusion results, but the Y-shape architecture can separately capitalize useful information from different modalities to generate more informative fused images.
- *SSIM loss*. Considering that simple pixel loss has low tolerance for noise, SSIM loss can avoid this defect and guide the network to obtain fusion results with more structural features.
- *SF loss*. With only SSIM loss, it is difficult to fully explore texture detail. Therefore, combined with SF loss, fused images with clearer scenes can be generated.
- *Unsupervised*. Due to the lack of ground-truth for infrared and visible image fusion, some methods train the fusion models with a specific database in a supervised manner, which is not appropriate for infrared and visible image fusion task. Thus, an unsupervised training strategy is more proper for multi-modality image fusion.
- *Generalization Ability*. Generalizing the trained deep fusion model to other image fusion tasks without fine-tuning plays an important role in identifying the practical application value. Nevertheless, most of the existing image fusion methods are designed for specific image fusion tasks and need to retrain the deep model or reset some parameters to handle other datasets. To this end, these methods can be regarded as lack of generalization ability.

## III. METHOD

In this section, the framework of the proposed YDTR is first introduced in Section III-A. Then, the network architecture of the dynamic Transformer module (DTRM) is described in detail in Section III-B. Finally, the loss function designed in this work is presented in Section III-C.

### A. Framework Overview

The network structure of the proposed YDTR is illustrated in Fig. 2, which consists of two Y branches and a main path. Because the infrared and visible image fusion aims to generate a composite image that exhibits salient targets with abundant scene details, the infrared image $I_{ir}$ and visible image $I_{vi}$ are separately fed into the two Y branches to capture the thermal radiation information from $I_{ir}$ and texture details from $I_{vi}$, respectively. Since combining CNN and Transformer has better feature exploration ability than pure Transformer [51], each Y branch is composed of an encoder and a DTRM. The encoder is designed for shallow local feature preservation and sequentially involves a convolutional layer with a kernel size of 3, a batch normalization (BN) layer, and a rectified linear unit (ReLU). The DTRM is devised to model high-level global dependencies, the network architecture of which is provided in Section III-B. Equipped with these two Y branches, more useful information from the source images can be adequately maintained. Then, the features extracted by the two Y branches are added together and fed into the main path. The main path includes a DTRM and a decoder. The DTRM is designed to merge the extracted features in depth. The decoder is employed for dimension reduction and fusion result generation, which contains a convolutional layer with a kernel size of 1 and a tanh activation function. Since the proposed YDTR is an end-to-end network, the output of the network is the fused image $I_f$.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

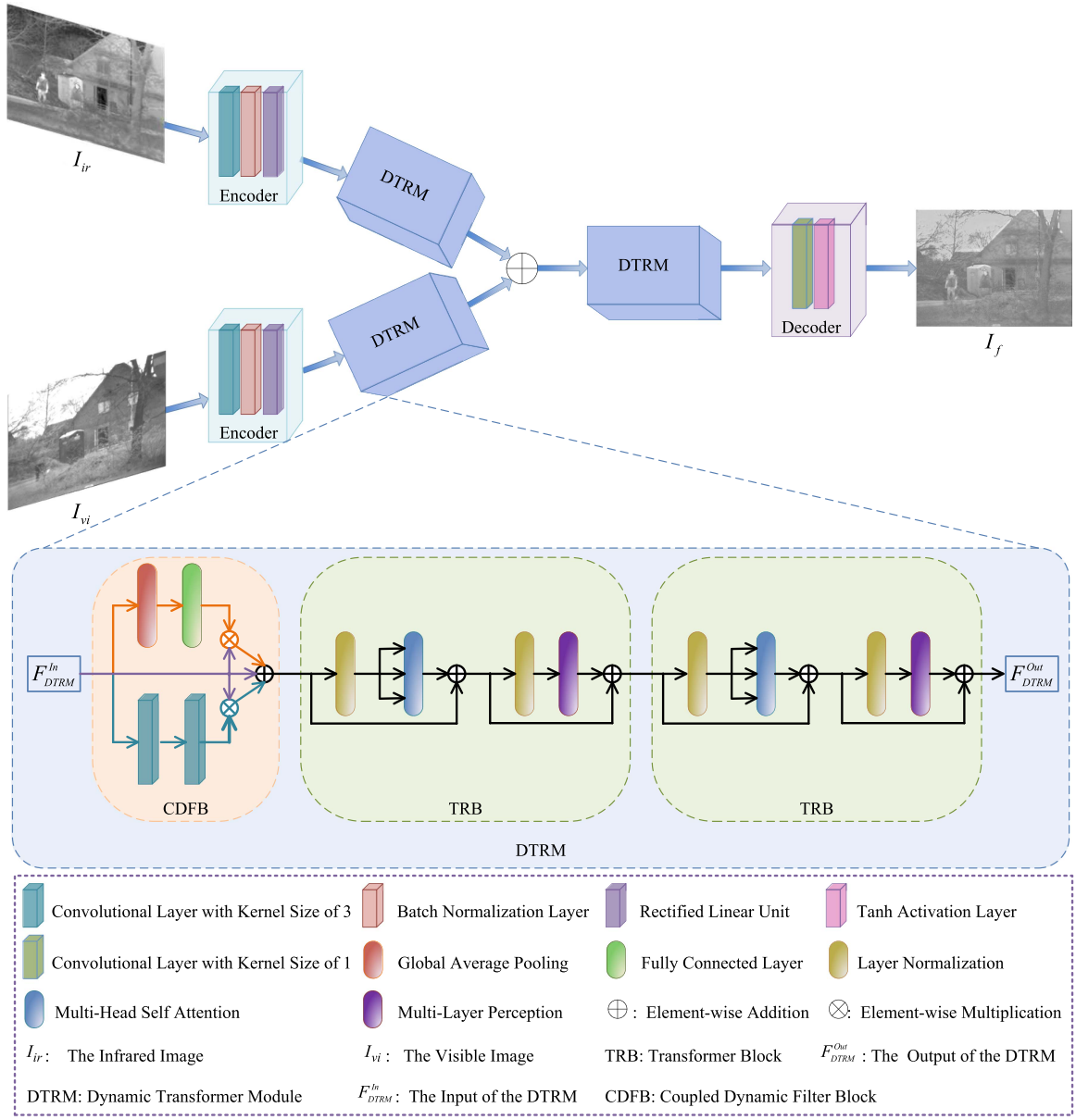


Fig. 2. The framework of the proposed YDTR for infrared and visible image fusion. The proposed Y-shape network is composed of two Y branches and a main path. First, the infrared image $I_{ir}$ and the visible image $I_{vi}$ are fed into the two Y branches, respectively. Each branch consists of an encoder and a DTRM. Afterwards, these two branches are added and fed into the main path, which involves a DTRM and a decoder, to obtain the fusion result $I_f$.

### *B. Dynamic Transformer Module*

As the aim of infrared and visible image fusion is to generate a composite image that can provide salient targets with abundant texture scene details. How to fully exploit complementary information from the source images is a crucial element that determines the fusion performance. With this consideration, we propose DTRM. The network structure of the DTRM is revealed in the blue block of Fig. 2. To comprehensively incorporate the explored features, the input of the DTRM $F^{In}_{DTRM}$ is first fed into the coupled dynamic filter block (CDFB), which can be formulated as

$$F_{CDFB} = F_{CF} \cdot F^{In}_{DTRM} + F^{In}_{DTRM} + F^{In}_{DTRM} \cdot F_{SF}, \quad (1)$$

where $F_{CDFB}$ means the features filtered by the CDFB. $F_{CF}$ represents the features filtered by the channel filter, which is deployed to encode channel-specific features and composed of a global average pooling (GAP) and a fully connected (FC) layer. $F_{SF}$ stands for the features filtered by the spatial filter, which is employed to implement filtering on account of semantic contents and is constituted of two $3 \times 3$ convolutional layers. Instead of extracting features through a convolution with a fixed kernel size, the proposed channel filter and spatial filter can simultaneously and dynamically preserve features. To this end, more complementary information can be content-adaptive maintained so as to obtain more informative fusion results. Then, $F_{CDFB}$ is fed into two successive Transformer blocks (TRBs) to further merge the complementary information with the consideration of long-term

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

relationships. Given the input of TRB $F_{CDFB}$ with the size of $h \times w \times c$, TRB first reshapes $F_{CDFB}$ to a $\frac{hw}{n^2} \times n^2 \times c$ feature by partitioning $F_{CDFB}$ into non-overlapping $n \times n$ local windows, where $\frac{hw}{n^2}$ represents the total number of windows. Afterwards, standard self-attention is separately performed on each window. For a local window feature $F_w \in \mathbb{R}^{n^2 \times c}$, the query, key, and value: $Q$, $K$, and $V$ are expressed as

$$M = XP_M(M = Q, K, V), \tag{2}$$

where $P_M$ stands for projection matrices that are shared across different windows. The attention matrix is calculated as

$$Attention(M) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{3}$$

where $B$ is the learnable relative positional encoding. Each TRB has two addition operations, and the first operation can be formulated as

$$F^1_{\text{add}} = F_{CDFB} + MSA\left(LN\left(F_{CDFB}\right)\right), \tag{4}$$

where $F^1_{\text{add}}$ denotes the output of the first addition operation in TRB. $MSA(\cdot)$ and $LN(\cdot)$ represent multi-head self-attention and layer normalization, respectively. The second addition operation can be formulated as

$$F^2_{add} = F^1_{add} + MLP\left(LN\left(F^1_{add}\right)\right), \tag{5}$$

where $F^2_{\text{add}}$ means the output of the second addition operation in TRB, and $MLP(\cdot)$ indicates the multi-layer perception.

### C. Loss Function

As the proposed YDTR is trained in an unsupervised manner, the loss function plays an important role in determining the fusion performance. Therefore, how to comprehensively take the distinct characteristics of the source images, e.g., the thermal radiation information with low spatial resolution of the infrared image and scene details with inconspicuous targets of the visible image, into consideration is critical. In this work, to fully preserve the complementary information from the source images, we design a loss function that contains two terms, which is formulated as

$$L = L_{SSIM}\left(I_f, I_s\right) + L_{SF}\left(I_f, I_s\right), \tag{6}$$

where $I_s$ stands for the source images. $L_{\text{SSIM}}(I_f, I_s)$ and $L_{\text{SF}}(I_f, I_s)$ represent the structural term and spatial frequency term, respectively.

Considering that simple pixel loss has low tolerance for noise, $L_{\text{SSIM}}(I_f, I_s)$ is devised to avoid this defect and ensure that the fused image has similar structural information to the source images, which can be formulated as

$$L_{SSIM}\left(I_f, I_s\right) = 1 - SSIM\left(I_f, I_s\right), \tag{7}$$

where $SSIM(I_f, I_s)$ is the structural similarity operation [52] between the fusion result and the source images, which is formulated as

$$SSIM\left(I_f, I_s\right) = \frac{\left(2\mu_s\mu_f + C_1\right)\left(2\sigma_{sf} + C_2\right)}{\left(\mu_s^2 + \mu_f^2 + C_1\right)\left(\sigma_s^2 + \sigma_f^2 + C_2\right)}, \tag{8}$$

where $C_1$ and $C_2$ represent constants devised to avoide instability when $\mu_s^2 + \mu_f^2$ or $\sigma_s^2 + \sigma_f^2$ is close to 0. $\mu_s$ and $\mu_f$ are the average intensities of the source images and fusion result, severally. $\sigma_{sf}$ stands for the covariance of $I_s$ and $I_f$. $\sigma_s^2$ and $\sigma_f^2$ denote the variances of $I_s$ and $I_f$, respectively.

With only $L_{\text{SSIM}}(I_f, I_s)$, it is difficult to fully explore the texture details. Thus, $L_{SF}(I_f, I_s)$ is employed to force the fusion result containing clear texture details, and is defined as

$$L_{SF}\left(I_f, I_s\right) = \left\|SF\left(I_f\right) - SF\left(I_s\right)\right\|_2, \tag{9}$$

where $SF(I_z), (z = f, s)$ means the spatial frequency (SF) operation [53] between input and output images. $\|\cdot\|_2$ denotes the $L_2-$Norm. SF is calculated by horizontal and vertical gradients, which reflects the change in image gray level and is formulated as

$$SF = 1 - \sqrt{Hor^2 + Ver^2}, \tag{10}$$

where $Hor$ and $Ver$ represent the horizontal and vertical gradients, respectively. The horizontal gradient is calculated by

$$Hor = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=2}^{W}\left|I(i,j) - I(i,j-1)\right|^2}, \tag{11}$$

where $H$ and $W$ denote the height and weight of an image $I$, respectively. The vertical gradient is defined as

$$Ver = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=2}^{W}\left|I(i,j) - I(i-1,j)\right|^2}. \tag{12}$$

Therefore, the loss function can be rewritten as

$$\begin{aligned} L = {} & L_{SSIM}\left(I_f, I_{ir}\right) + \alpha \cdot L_{SSIM}\left(I_f, I_{vi}\right) \\ & + \beta \cdot L_{SF}\left(I_f, I_{ir}\right) + \gamma \cdot L_{SF}\left(I_f, I_{vi}\right), \end{aligned} \tag{13}$$

where $\alpha$, $\beta$, and $\gamma$ are three trade-off parameters that control the balance of the loss function. The impacts of these three factors are analyzed in Section IV-C.

## IV. Experiments

In this section, the datasets employed in this work and experimental details are first described in Section IV-A. Then, Section IV-B introduces the compared approaches and objective evaluation metrics. The ablation analyses on the loss function and network structure are provided in Section IV-C. The experimental results and discussion are reported in Section IV-D. Generalization experiments on infrared and RGB-visible dataset are presented in Section IV-E. Finally, we further extend the proposed method to address the multi-focus image fusion task in Section IV-F.

### A. Datasets and Experimental Details

Two mainstream databases are utilized in this work: the TNO database[1] and RoadScene database [20]. In total, 348 infrared and visible image pairs are collected from these two datasets

[1]https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

and randomly divided into a training set, a validation set, and a testing set with 288, 20, and 40 image pairs, respectively. To acquire sufficient training samples, the over-lapping cropping strategy is utilized to augment the training set. It is worth mentioning that the cropping strategy is a widely used method for data augmentation in the field of image fusion [2], [20], [54]–[57]. Concretely, each image is cropped into $120 \times 120$ with a stride of 20, therefore, 58708 pairs of infrared and visible image patches are generated for network training. All training samples are normalized to [0,1]. Since the cropping strategy is only employed for data augmentation, it is not adopted for the validation set or the testing set. Therefore, the fusion results can be generated through feeding the entire image into the trained model.

In our experiments, the epoch is 10 and the batch size is fixed as 32. The learning rate is set to 0.001 and the Adam optimizer [58] is adopted for model optimization. The three weight factors $\alpha$, $\beta$, and $\gamma$ in the loss function are specified as 0.05, 0.0006, and 0.00025, respectively. All experiments are conducted on a computer with an NVIDIA GeForce RTX 3090 GPU and the proposed deep model is implemented on the PyTorch framework.

### B. Comparison Methods and Fusion Metrics

Nine infrared and visible image fusion methods are adopted to compare the fusion performance, including two representative conventional methods, i.e., the ratio Pyramid (RP)-based method [8] and curvelet transform (CVT)-based method [9], and seven state-of-the-art DL-based methods, e.g., the DenseFuse-based method [18], FusionGAN-based method [19], U2Fusion-based method [20], GANMcC-based method [24], RFN-Nest-based method [23], CSF-based method [29], and PPT Fusion-based method [32]. The codes of all nine algorithms are publicly available, and all parameters are set as the default values as their original publications reported for unbiased comparison.

As reported in [59], image fusion evaluation metrics can be classified into four categories: information theory-based metrics, image feature-based metrics, image structural similarity-based metrics, and human perception inspired metrics. With this into consideration, in this paper, five widely applied evaluation metrics are deployed for comprehensive quantitative comparison, which include the normalized mutual information $Q_{MI}$ [46], nonlinear correlation information entropy $Q_{NCIE}$ [47], phase congruency based feature-wise metric $Q_P$ [48], multi-scale structural similarity index ($MS\text{-}SSIM$) [49], and Chen-Varshney metric $Q_{CV}$ [50]. Due to the aim of image fusion is to generate an informative composite image, two information theory-based metrics are utilized, i.e., $Q_{MI}$ and $Q_{NCIE}$. $Q_{MI}$ assesses the information that is transformed from the source images to the fusion image. $Q_{NCIE}$ calculates the nonlinear correlation information entropy among the fusion and the source images. $Q_P$ is an image feature-based metric that evaluates the fusion performance through contradistinguishing the local correlation between input and fusion images in terms of feature maps. $MS\text{-}SSIM$ is an image structural similarity-based metric. It calculates the structural similarity between the input and the fusion images. $Q_{CV}$ is a human perception-inspired fusion metric, which is consistent with the human visual system (HVS). For $Q_{MI}$, $Q_{NCIE}$, $Q_P$, and $MS\text{-}SSIM$, a higher evaluation value indicates better fusion performance, while for $Q_{CV}$, a lower score indicates a better fusion quality.

TABLE II
OBJECTIVE EVALUATION RESULTS OF THE PROPOSED METHOD WITH DIFFERENT LOSS FUNCTIONS

| Metrics | $L_1$ | $L_2$ | $L_3$ | $L_{total}$ |
|---|---|---|---|---|
| $Q_{MI}$ [46] ↑ | 0.1237 | 0.3240 | 0.4013 | **0.4401** |
| $Q_{NCIE}$ [47] ↑ | 0.8032 | 0.8055 | 0.8065 | **0.8078** |
| $Q_P$ [48] ↑ | 0.0514 | 0.2367 | 0.3535 | **0.3938** |
| $MS\text{-}SSIM$ [49] ↑ | 0.4726 | 0.6698 | 0.6702 | **0.6884** |
| $Q_{CV}$ [50] ↓ | 96.7566 | 92.992 | 75.6981 | **69.2637** |

### C. Ablation Study

To illustrate the significance of the proposed method, we implement ablation studies on two important components to investigate their influence: the network structure and the loss function. All ablation analyses are conducted on the validation set.

*1) Ablation Analysis on the SSIM Term and SF Term in the Loss Function:* Since the proposed YDTR is an end-to-end unsupervised fusion model, the loss function plays a crucial role in affecting the fusion performance. With this into consideration, we design a loss function composed of two terms, i.e., $L_{SSIM}$ and $L_{SF}$, to guide the proposed fusion model to fully extract complementary information from the source images. In order to verify the significance of these two terms, ablation experiments are conducted. Specifically, to investigate the effectiveness of $L_{SSIM}$, we discard this term from the loss function. To this end, $L_{total}$ is reformulated as

$$L_1 = L_{SF}\left(I_f, I_s\right). \tag{14}$$

Similarly, we remove $L_{SF}$ from $L_{total}$ to demonstrate its necessity, and rewrite $L_{total}$ as

$$L_2 = L_{SSIM}\left(I_f, I_s\right). \tag{15}$$

To further illustrate the irreplaceability of the proposed $L_{SF}$, we replace $L_{SF}$ with a gradient operator, and redefine $L_{total}$ as

$$L_3 = L_{SSIM}\left(I_f, I_s\right) + L_G\left(I_f, I_s\right), \tag{16}$$

where $L_G(I_f, I_s)$ is the gradient term, which is expressed as

$$L_G\left(I_f, I_s\right) = \left\|\nabla I_f - \nabla I_s\right\|_2, \tag{17}$$

where $\nabla$ represents the gradient operation.

The quantitative assessment results of the proposed method with different loss functions are listed in Table II. For each metric, the average score of all validation samples is reported. It is obvious that without $L_{SSIM}$ or $L_{SF}$, the objective evaluation values on $Q_{MI}$, $Q_{NCIE}$, $Q_P$, and $MS\text{-}SSIM$ are largely decreased, while the scores on $Q_{CV}$ are greatly increased, which demonstrates that with only $L_{SSIM}$ or $L_{SF}$, satisfactory fusion results cannot be generated. Compared with $L_1$ and $L_2$, $L_3$ reveals better objective performance, but a wide gap remains when

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

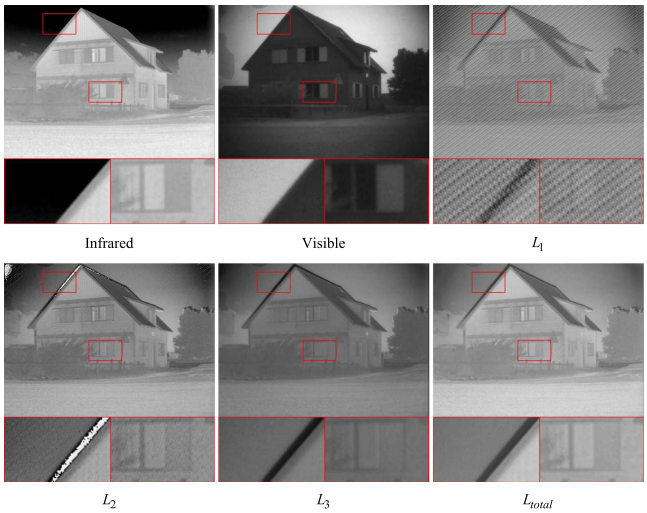


Fig. 3. A pair of source images and their corresponding fusion results of the proposed method with different loss functions.

compared with $L_{total}$, which further illustrates the significance of $L_{SF}$.

Fig. 3 reports a pair of source images and their corresponding fusion results generated by different loss functions. Similar observations can be obtained with the quantitative performance. Specifically, undesirable noise occurs on $L_1$, which severely degrades the fusion result. $L_2$ has better performance than $L_1$ but still suffers from rebarbative artifacts. $L_3$ eliminates noise while losing some thermal radiation information, leading to indistinct target (see the window in Fig. 3). Overall, $L_{total}$ owns the best subjective performance in terms of simultaneously preserving complementary information.

*2) Ablation Analysis on the Weight Parameters in the Loss Function:* In the proposed method, three trade-off parameters, i.e., $\alpha$, $\beta$, and $\gamma$, control the contribution of the loss function, which are fixed as 0.05, 0.0006, and 0.00025 on the basis of a large number of experiments. Since there exist too many combinations of these three parameters, one set of fusion performance is reported to demonstrate the impact of each factor by specifying the other two as the default scores.

To investigate the impact of $\alpha$ on the fusion quality, $\beta$ and $\gamma$ are set to 0.0006 and 0.00025, respectively. Through changing $\alpha$ (i.e., 0.0005, 0.005, 0.5, and 5), the influence on the fusion results can be observed. A pair of infrared and visible images and the fusion results generated by $\alpha$ with different values are shown in Fig. 4. It can be found that when $\alpha$ decreases (0.0005 and 0.005), the fused images contain fewer texture details of the visible image (see the first close-ups in Fig. 4). When $\alpha$ increases (0.5, and 5), the fusion results preserve less thermal radiation information from the infrared image (see the second close-ups in Fig. 4). Overall, when $\alpha$ is fixed as 0.05, the fusion result owns appropriate complementary information preservation from the source images. Table III reports quantitative assessments of the proposed method using different values of $\alpha$ in the loss function ($\beta$ and $\gamma$ are fixed as 0.0006 and 0.00025, respectively). For each metric, the average value of all validation samples is listed, and the best one is labeled in bold. It can be observed that when

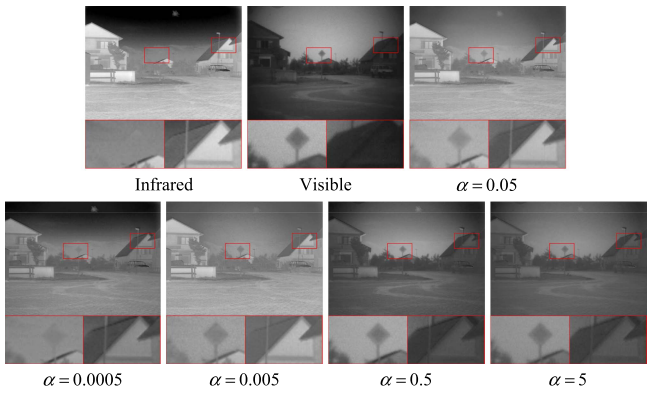


Fig. 4. A pair of source images and their corresponding fusion results of the proposed method using different values of $\alpha$ in the loss function ($\beta$ and $\gamma$ are fixed as 0.0006 and 0.00025, respectively).

TABLE III
OBJECTIVE EVALUATION OF THE PROPOSED METHOD USING DIFFERENT VALUES OF $\alpha$ IN THE LOSS FUNCTION ($\beta$ AND $\gamma$ ARE FIXED AS 0.0006 AND 0.00025, RESPECTIVELY)

| Metrics | 0.0005 | 0.005 | 0.05 | 0.5 | 5 |
|---|---|---|---|---|---|
| $Q_{MI}$ [46] ↑ | 0.4148 | 0.4063 | **0.4401** | 0.4338 | 0.3939 |
| $Q_{NCIE}$ [47] ↑ | 0.8075 | 0.8074 | **0.8078** | 0.8037 | 0.8067 |
| $Q_P$ [48] ↑ | 0.3623 | 0.3542 | **0.3938** | 0.3641 | 0.3096 |
| $MS\text{-}SSIM$ [49] ↑ | 0.6822 | 0.6652 | **0.6884** | 0.6634 | 0.6574 |
| $Q_{CV}$ [50] ↓ | 127.0317 | 72.6817 | **69.2637** | 167.9059 | 112.4705 |

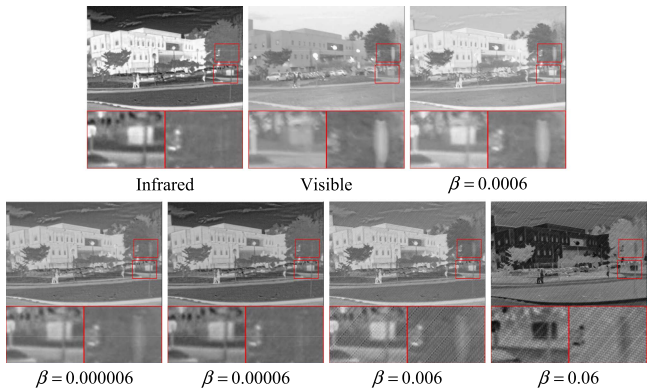


Fig. 5. A pair of source images and their corresponding fusion results of the proposed method using different values of $\beta$ in the loss function ($\alpha$ and $\gamma$ are fixed as 0.05 and 0.00025, respectively).

$\alpha$ is set to 0.05, it can achieve the best objective performance. Increasing or decreasing $\alpha$, worse objective performance will be obtained. Based on the above subjective and objective analysis, we set $\alpha$ as 0.05.

Similarly, $\alpha$ and $\gamma$ are specified as 0.05 and 0.00025 to demonstrate the influence of $\beta$ (i.e., 0.000006, 0.00006, 0.006 and 0.06) on the fusion results. Fig. 5 exhibits a pair of source images and the corresponding fused images. Obviously, when $\beta$ is fixed as 0.0006, the fusion result achieves the best performance. Specifically, it reveals salient targets (see the first close-ups in Fig. 5) and clear details (see the second close-ups in Fig. 5). However, when $\beta$ increases (0.006 and 0.06), unexpected noise occurs; when $\beta$ decreases (0.000006 and 0.00006), the fusion results

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
OBJECTIVE EVALUATION OF THE PROPOSED METHOD USING DIFFERENT VALUES OF $\beta$ IN THE LOSS FUNCTION ($\alpha$ AND $\gamma$ ARE FIXED AS 0.05 AND 0.00025, RESPECTIVELY)

| Metrics | 0.00006 | 0.00006 | 0.0006 | 0.006 | 0.06 |
|---|---|---|---|---|---|
| $Q_{MI}$ [46] ↑ | 0.4267 | 0.4329 | **0.4401** | 0.194 | 0.1764 |
| $Q_{NCIE}$ [47] ↑ | 0.8069 | 0.8024 | **0.8078** | 0.8039 | 0.8027 |
| $Q_P$ [48] ↑ | 0.344 | 0.3199 | **0.3938** | 0.1541 | 0.0764 |
| $MS\text{-}SSIM$ [49] ↑ | 0.6733 | 0.6512 | **0.6884** | 0.6005 | 0.5996 |
| $Q_{CV}$ [50] ↓ | 164.2465 | 179.7848 | **69.2637** | 115.9945 | 364.4186 |

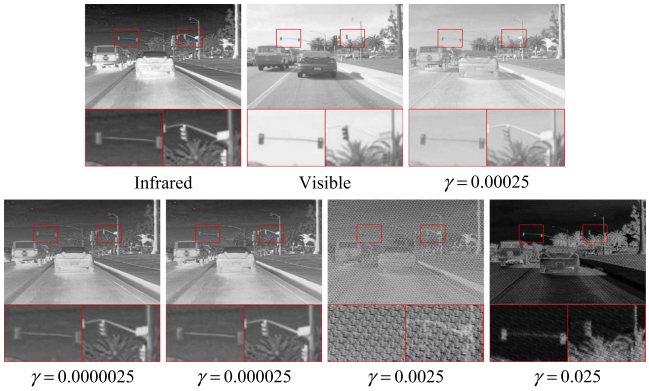


Fig. 6. A pair of source images and their corresponding fusion results of the proposed method using different values of $\gamma$ in the loss function ($\alpha$ and $\beta$ are fixed as 0.05 and 0.0006, respectively).

TABLE V
OBJECTIVE EVALUATION OF THE PROPOSED METHOD USING DIFFERENT VALUES OF $\gamma$ IN THE LOSS FUNCTION ($\alpha$ AND $\beta$ ARE FIXED AS 0.05 AND 0.0006, RESPECTIVELY)

| Metrics | 0.000025 | 0.000025 | 0.00025 | 0.0025 | 0.025 |
|---|---|---|---|---|---|
| $Q_{MI}$ [46] ↑ | 0.4344 | 0.4248 | **0.4401** | 0.11 | 0.2443 |
| $Q_{NCIE}$ [47] ↑ | 0.8043 | 0.8041 | **0.8078** | 0.8032 | 0.8034 |
| $Q_P$ [48] ↑ | 0.374 | 0.3385 | **0.3938** | 0.0504 | 0.1068 |
| $MS\text{-}SSIM$ [49] ↑ | 0.6734 | 0.6671 | **0.6884** | 0.39 | 0.3206 |
| $Q_{CV}$ [50] ↓ | 166.6168 | 198.5401 | **69.2637** | 332.6176 | 462.2035 |

can neither well preserve infrared information nor texture details. Table IV shows the quantitative evaluation of the proposed method using different values of $\beta$ in the loss function ($\alpha$ and $\gamma$ are fixed as 0.05 and 0.00025, respectively). For each metric, the average score of all validation samples is listed, and the best one is labeled in bold. Obviously, the best objective fusion performance is reached when $\beta$ is set to 0.0006.

To characterize the influence of $\gamma$ (viz. 0.0000025, 0.000025, 0.0025, and 0.025), $\alpha$ and $\beta$ are fixed as 0.05 and 0.0006, severally. Fig. 6 provides a set of source and fusion images. Some undesirable noise exists in the fused results when $\gamma$ is set to 0.0025 or 0.025. When $\gamma$ is 0.0000025 or 0.000025, the fusion results possess few scene details. On the whole, high fusion quality is acquired when $\gamma$ is specified as 0.00025. In this circumstance, both thermal radiation information of the infrared image and texture details of the visible image are well preserved in the fusion result. Table V reports the objective evaluation of the proposed method using different values of $\gamma$ in the loss function ($\alpha$ and $\beta$ are fixed as 0.05 and 0.0006, respectively). For each metric, the average value of all validation images is shown, and the optimal one is listed in bold. Clearly, when $\gamma$ is set to 0.00025, the best fusion performance can be obtained.

*3) Ablation Analysis on the Network Architecture:* The proposed network structure has two main characteristics, namely, the Y-shape network and the DTRM. The Y-shape network is designed to separately capture the complementary information from the source images through the two Y branches and adequately integrate the extracted features by the main path. The DTRM, which is composed of CDFB and TRB, is devised to ensure that our network can not only capture local significant features but also model long-range dependencies. To investigate the effectiveness of these four components, i.e., the Y-shape architecture, DTRM, CDFB, and TRB, we conduct ablation experiments on the validation set and modify the proposed YDTR as follows. a) Single Path. To verify the effectiveness of the proposed Y shaped network, we modify the YDTR using a single path. In this way, the source images are first concatenated in the channel dimension to obtain a two-channel map $\{I_{ir}, I_{vi}\}$, then, $\{I_{ir}, I_{vi}\}$ is fed into a single path to generate the fusion result. b) Concat. In our proposed method, the features extracted by the two Y branches are added for feature aggregation. To demonstrate the superiority of the addition aggregation, we conduct an ablation study by concatenating the exploited features. c) w/o CDFB. In order to identify the significance of the CDFB, we remove it from the intact model to investigate its function. d) w/o TRB. To investigate the necessity of the TRB and demonstrate the significance of long-range relationship construction, we delete it from the YDTR to illustrate its effectiveness. e) w/o DTRM. To elucidate the importance of the DTRM, we delete all DTRMs from our proposed model and add encoders at the same position to have an identical structure to the YDTR. f) w/o specific DTRM. To further demonstrate the role of each DTRM, several ablation experiments are conducted. Specifically, the DTRM in the infrared branch, visible branch, and main path are named as $\mathrm{DTRM}_1$, $\mathrm{DTRM}_2$, and $\mathrm{DTRM}_3$, respectively. We replace $\mathrm{DTRM}_i$ ($i \in [1, 3]$) with an encoder to investigate the influence of removing one DTRM on the fusion performance. Similarly, we replace two DTRMs with encoders to inquire the impact of deleting two DTRMs on the fusion result.

Fig. 7 illustrates a pair of source images and their corresponding fusion results generated by different network architectures. In each image, two local areas are zoomed as close-ups for better comparison. It can be observed that Single Path loses some important texture details, resulting in unclear scene, which demonstrates that the proposed Y-shape network owns better complementary information preservation capability. Concat shows better detail extraction ability than Single Path, but it fails to adequately exploit thermal radiation information. Therefore, our addition aggregation has better feature integration competence. w/o CDFB can produce good fusion result, but it cannot fully maintain the significant features, leading to blurry scene and indistinctive object. Thus, equipped with CDFB, our model exhibits better feature extraction ability. Similarly, for w/o TRB, the fusion model cannot construct long-range dependencies well, so some texture details and targets are lost. Therefore,

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

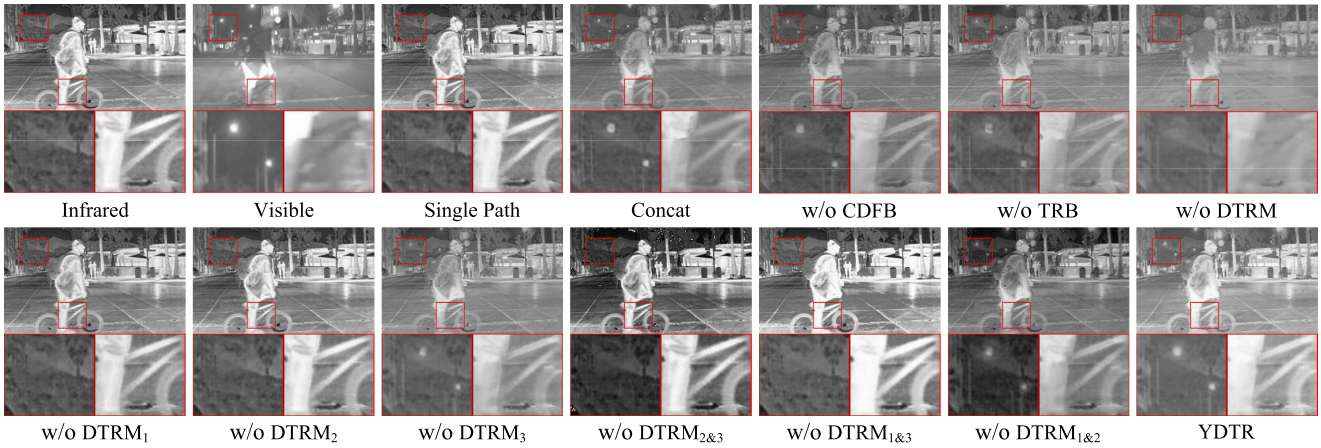


Fig. 7. A pair of source images and their corresponding fusion results of the proposed method with different network structures.

TABLE VI
OBJECTIVE EVALUATION RESULTS OF THE PROPOSED METHOD WITH DIFFERENT NETWORK STRUCTURES

| | $Q_{MI}$ [46] ↑ | $Q_{NCIE}$ [47] ↑ | $Q_P$ [48] ↑ | $MS\text{-}SSIM$ [49] ↑ | $Q_{CV}$ [50] ↓ |
|---|---|---|---|---|---|
| Single Path | 0.3733 | 0.8045 | 0.3017 | 0.6279 | 124.3958 |
| Concat | 0.3983 | 0.8067 | 0.329 | 0.6822 | 99.6134 |
| w/o CDFB | 0.3306 | 0.8054 | 0.2379 | 0.6747 | 102.5727 |
| w/o TRB | 0.3605 | 0.806 | 0.2507 | 0.6782 | 96.468 |
| w/o DTRM | 0.288 | 0.8046 | 0.1426 | 0.4665 | 126.1872 |
| w/o $\mathrm{DTRM}_1$ | 0.3586 | 0.8064 | 0.2352 | 0.4704 | 109.2994 |
| w/o $\mathrm{DTRM}_2$ | 0.3552 | 0.8063 | 0.2307 | 0.5377 | 105.1364 |
| w/o $\mathrm{DTRM}_3$ | 0.3707 | 0.8063 | 0.2681 | 0.6812 | 106.703 |
| w/o $\mathrm{DTRM}_{2\&3}$ | 0.2945 | 0.8057 | 0.2246 | 0.4697 | 118.25 |
| w/o $\mathrm{DTRM}_{1\&3}$ | 0.2943 | 0.806 | 0.2069 | 0.4693 | 119.6726 |
| w/o $\mathrm{DTRM}_{1\&2}$ | 0.3503 | 0.8059 | 0.2231 | 0.4696 | 120.6865 |
| YDTR | **0.4401** | **0.8078** | **0.3938** | **0.6884** | **69.2637** |

long-range interactions can facilitate vital complementary information preservation. For w/o DTRM, the fusion quality is much lower than YDTR. For one thing, the scene details cannot be well preserved. Specifically, the street lamp disappeared in the fusion result (see the first close-ups in Fig. 7). For another, the target is blurred, e.g., the man and bicycle are not salient (see the second close-ups in Fig. 7). Therefore, our proposed DTRM can improve the feature extraction ability of the fusion model in terms of both local useful information exploitation and global complementary feature preservation. For w/o one DTRM or two DTRMs, some useful information is lost, which demonstrates the necessity of the DTRM in the proposed fusion framework.

Table VI reports the objective assessment results of the proposed method with different network structures. For each metric, the average score of all validation samples is listed, and the optimal one is labeled in bold. Consistent with subjective observations, the intact model exhibits the best performance on all metrics.

### D. Experimental Results and Discussion

In this subsection, experimental results on two popular infrared and visible image fusion datasets are presented. We compare the proposed method with two representative traditional approaches and seven state-of-the-art algorithms in terms of both visual quality and objective evaluation indexes.

*1) Results on the TNO Dataset:* Fig. 8 reports four sets of source image pairs and their corresponding fusion results obtained by different methods. All nine compared methods perform well. However, compared with the YDTR, they still suffer from some defects. To be more specific, the RP-based method cannot well preserve useful information from the source images resulting in indistinct targets and blurry scenes. The CVT-based method is slightly unnatural. The DenseFuse-based and U2Fusion-based methods have limited abilities to capture the thermal radiation information from the infrared image. The texture details are lost in the FusionGAN-based and GANMcC-based methods. The RFN-Nest-based method darkens the entire image leading to lower contrast. The CSF-based and PPT Fusion-based methods cannot well maintain the thermal radiation information of the infrared image, leading to indistinctive target. Overall, the proposed YDTR achieves the best fusion results from the perspective of visual performance. On one hand, the proposed method can explore the major significant information from the infrared image revealing distinguished targets. On the other hand, sufficient texture details of the visible image are adequately maintained by the YDTR providing a clear scene background. Table VII reports the quantitative assessments of different fusion methods. The mean values of 20 image pairs from the TNO dataset are listed. For each metric, the best and second best fusion results of all methods are labeled in bold and underlined, respectively. It can be observed that our proposed

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
QUANTITATIVE EVALUATION RESULTS OF NINE REPRESENTATIVE AND STATE-OF-THE-ART INFRARED AND VISIBLE IMAGE FUSION METHODS AND THE PROPOSED YDTR ON TNO DATASET

| | $Q_{MI}$ [46] ↑ | $Q_{NCIE}$ [47] ↑ | $Q_P$ [48] ↑ | $MS\text{-}SSIM$ [49] ↑ | $Q_{CV}$ [50] ↓ |
|---|---|---|---|---|---|
| RP [8] | 0.2449 | 0.8038 | 0.2498 | 0.6400 | 61.1259 |
| CVT [9] | 0.2420 | 0.8037 | 0.2721 | 0.6572 | <u>45.7575</u> |
| DenseFuse [18] | <u>0.3390</u> | 0.8050 | <u>0.2936</u> | 0.6769 | 47.9642 |
| FusionGAN [19] | 0.3335 | <u>0.8050</u> | 0.1025 | 0.6123 | 98.5990 |
| U2Fusion [20] | 0.2936 | 0.8045 | 0.2815 | 0.6594 | 49.6913 |
| GANMcC [24] | 0.3314 | 0.8047 | 0.2359 | 0.6619 | 65.3504 |
| RFN-Nest [23] | 0.2979 | 0.8045 | 0.2506 | 0.6687 | 54.5542 |
| CSF [29] | 0.2956 | 0.8045 | 0.2561 | 0.6765 | 48.0904 |
| PPT Fusion [32] | 0.3162 | 0.8047 | 0.2691 | **0.6902** | 64.0136 |
| YDTR | **0.3526** | **0.8053** | **0.2978** | <u>0.6827</u> | **41.5507** |

The mean values of 20 image pairs from the TNO dataset are listed. For each metric, the optional and suboptional values are labeled in bold and underlined, Respectively.

TABLE VIII
QUANTITATIVE EVALUATION RESULTS OF NINE REPRESENTATIVE AND STATE-OF-THE-ART INFRARED AND VISIBLE IMAGE FUSION METHODS AND THE PROPOSED YDTR ON ROADSCENE DATA SET

| | $Q_{MI}$ [46] ↑ | $Q_{NCIE}$ [47] ↑ | $Q_P$ [48] ↑ | $MS\text{-}SSIM$ [49] ↑ | $Q_{CV}$ [50] ↓ |
|---|---|---|---|---|---|
| RP [8] | 0.3035 | 0.8054 | 0.3322 | 0.6128 | 139.9316 |
| CVT [9] | 0.2873 | 0.8051 | 0.3119 | 0.6282 | 116.8517 |
| DenseFuse [18] | 0.3943 | 0.8070 | 0.3520 | **0.6712** | 95.0244 |
| FusionGAN [19] | 0.3800 | 0.8068 | 0.1220 | 0.5692 | 138.8463 |
| U2Fusion [20] | 0.3403 | 0.8059 | 0.3349 | 0.6597 | 96.0016 |
| GANMcC [24] | 0.3620 | 0.8062 | 0.2791 | 0.6373 | 111.9141 |
| RFN-Nest [23] | 0.3618 | 0.8063 | 0.2521 | 0.6585 | 119.4961 |
| CSF [29] | 0.3668 | 0.8065 | 0.3356 | 0.6596 | <u>92.3366</u> |
| PPT Fusion [32] | <u>0.4030</u> | **0.8072** | <u>0.3546</u> | <u>0.6682</u> | 128.9598 |
| YDTR | **0.4035** | <u>0.8070</u> | **0.3592** | 0.6653 | **60.8876** |

The mean values of 20 image pairs from the roadscene data set are listed. For each metric, the optional and suboptional values are labeled in bold and underlined, respectively.

YDTR has obvious superiority over other competitors on $Q_{MI}$, $Q_{NCIE}$, $Q_P$, and $Q_{CV}$. For $MS\text{-}SSIM$, the proposed method has a suboptimal performance, and the margin to the best one is slight. Based on the above analysis, our method achieves satisfactory fusion performance in both subjective and objective evaluations.

*2) Results on the RoadScene Dataset:* Four groups of infrared and visible images on the RoadScene dataset and their corresponding fusion results obtained by the nine compared methods and proposed YDTR are shown in Fig. 9. Similar to the fusion performance on the TNO dataset, the proposed method owns the best visual fusion quality compared to other competitors. Specifically, the fusion results generated by the proposed YDTR reveal abundant scene details and distinct foreground objects. The texture details are lost to some degree in the RP-based and CVT-based methods. The FusionGAN-based method owns lower contrast, resulting in burred details. The U2Fusion-based method can well preserve the texture details from the visible image while failing to extract some typical features from the infrared image. The DenseFuse-based, GANMcC-based, and RFN-Nest-based methods reveal salient targets, but the scene details are lost to some extent. The CSF-based method loses some important texture details, resulting in blurry scene. The PPT Fusion-based method tends to under-preserve the thermal radiation information of the infrared image, leading to blurred objects. The objective evaluation results are listed in Table VIII. For each metric, the average value of all testing samples is calculated. The optimal and suboptimal scores are labeled in bold and underlined, respectively. It is shown that our YDTR obtains the optional values on $Q_{MI}$, $Q_P$, and $Q_{CV}$ with significant margins. For $Q_{NCIE}$ and $MS\text{-}SSIM$, the proposed method has relatively lower scores when compared with the latest approaches with a tiny gap. Overall, the proposed YDTR has a satisfactory objective evaluation performance. On the basis of the fusion performance on both qualitative and quantitative assessments, the proposed method reveals superiority over other algorithms.

TABLE IX
RUNNING TIME OF DIFFERENT METHODS (UNIT: SECONDS)

| Method | TNO Dataset | RoadScene Dataset |
|---|---|---|
| RP [8] | **0.0742** | **0.0483** |
| CVT [9] | 0.7086 | 0.4167 |
| DenseFuse [18] | 0.5663 | 0.3190 |
| FusionGAN [19] | 2.6796 | 1.1442 |
| U2Fusion [20] | 2.2085 | 1.1639 |
| GANMcC [24] | 5.6752 | 2.3813 |
| RFN-Nest [23] | 2.3096 | 0.9423 |
| CSF [29] | 10.3311 | 5.5395 |
| PPT Fusion [32] | 0.4126 | 0.2203 |
| YDTR | <u>0.1284</u> | <u>0.1626</u> |

*3) Efficiency Study:* To further conduct a comprehensive comparison with other state-of-the-art methods, we compare our YDTR with other approaches in the aspects of running time, parameters, and FLOPs. Table IX reports the running time of

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

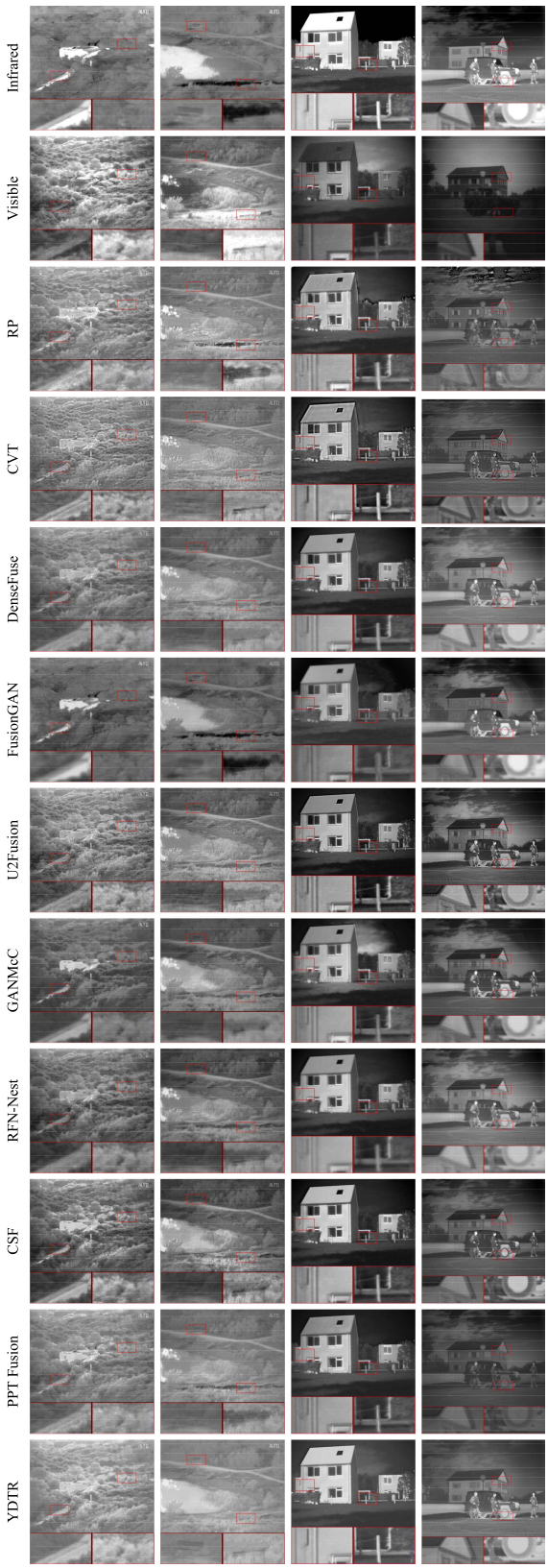


Fig. 8. Four sets of source images and their corresponding fusion results of different methods on TNO data set. From top to bottom: infrared image, visible image, the fusion results of RP [8], CVT [9], DenseFuse [18], FusionGAN [19], U2Fusion [20], GANMcC [24], RFN-Nest [23], CSF [29], PPT Fusion [32], and the proposed YDTR. In each image, two local areas are enlarged as close-ups for better comparison.

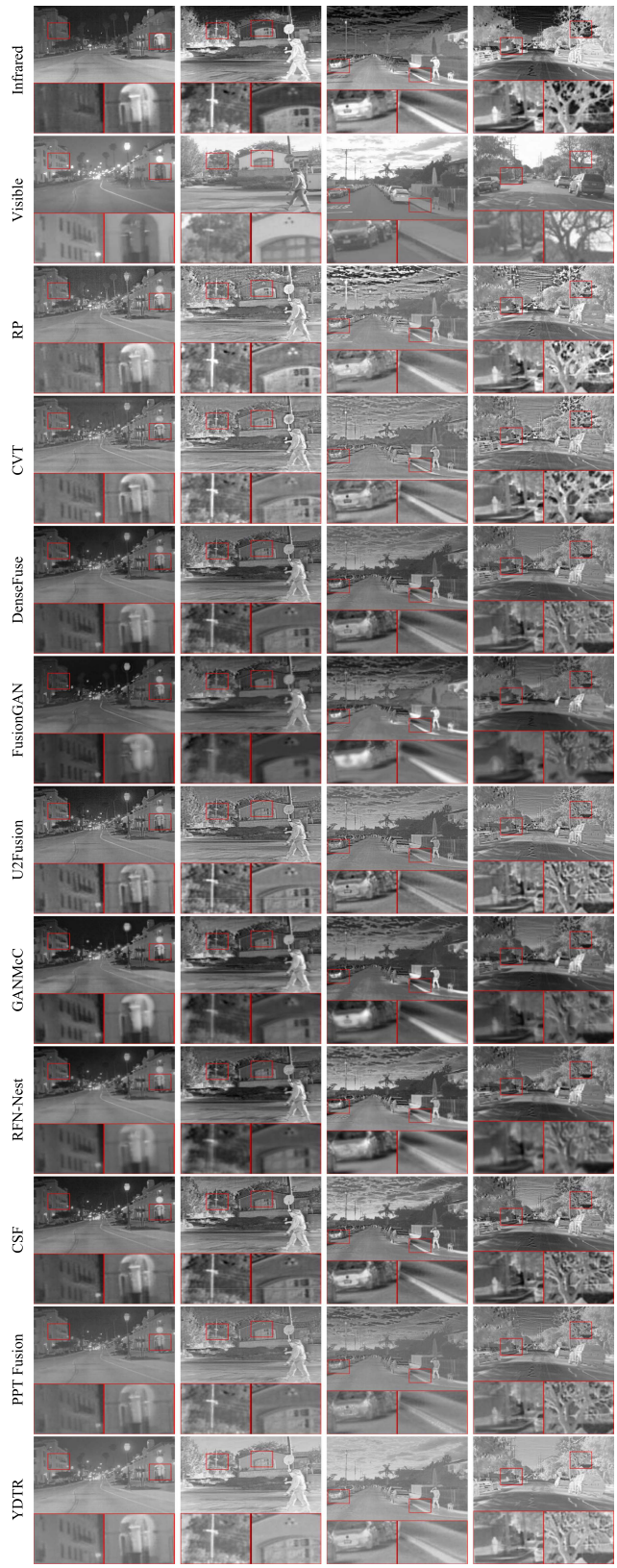


Fig. 9. Four sets of source images and their corresponding fusion results of different methods on RoadScene data set. From top to bottom: infrared image, visible image, the fusion results of RP [8], CVT [9], DenseFuse [18], FusionGAN [19], U2Fusion [20], GANMcC [24], RFN-Nest [23], CSF [29], PPT Fusion [32], and the proposed YDTR. In each image, two local areas are enlarged as close-ups for better comparison.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

TABLE X
EFFICIENCY COMPARISON ON PARAMETERS AND FLOPS

| Method | Paramters (M) | FLOPs (G) |
|---|---|---|
| DenseFuse [18] | <u>0.8903</u> | 6.7291 |
| FusionGAN [19] | 5.3056 | 0.5469 |
| U2Fusion [20] | 2.6369 | 7.8948 |
| GANMcC [24] | 9.1022 | 1.0237 |
| RFN-Nest [23] | 19.1660 | 7.6763 |
| CSF [29] | 1.6397 | **0.0004** |
| PPT Fusion [32] | 312.5742 | 52.0814 |
| YDTR | **0.8711** | <u>0.2113</u> |

different methods on two datasets. For each dataset, the average time of generating a fused image on all testing samples is calculated. The best and second-best methods are labeled in bold and underlined, respectively. It can be observed that the proposed method has obvious superiority compared with other deep fusion models. To investigate the computational complexity of different approaches, a $120 \times 120$ image is adopted. Table X lists the efficiency comparison on Parameters and FLOPs. Overall, the proposed YDTR owns obvious efficiency superiority.

### *E. Extension to Infrared and RGB-Visible Image Fusion*

To verify the generalization ability of the proposed method, we extend our YDTR to deal with infrared and RGB-visible image fusion issue without fine-tuning. The Multi-Spectral Road Scenarios (MSRS) dataset [60] is employed to conduct the generalization experiment. Specifically, 361 pairs of infrared and RGB-visible images are downloaded as testing samples. Given that RGB-visible images are three-channel data and infrared images are single-channel data, RGB-to-YUV color conversion is first implemented to get the $Y$, $U$, and $V$ components: $I_{vi}^{Y}$, $I_{vi}^{U}$, and $I_{vi}^{V}$. It is worth mentioning that the RGB-to-YUV color conversion is widely used to address the channel mismatching problem [54], [55], [61]. Then, $I_{vi}^{Y}$ and infrared image $I_{ir}$ are fed into our trained model to generate the $Y$ component of the fusion result $I_{f}^{Y}$. Finally, YUV-to-RGB color conversion among $I_{f}^{Y}$, $I_{vi}^{U}$, and $I_{vi}^{V}$ is performed to obtain the final fused image $I_{f}$.

Fig. 10 shows four sets of source images and their corresponding fusion results acquired by nine state-of-the art methods and our proposed YDTR. In each image, two local areas are enlarged as close-ups for better comparison. All ten approaches can produce good fusion results. However, compared with YDTR, the other competitors still remain several drawbacks. It can be observed that the RP-based method suffers from severe noise. The CVT-based method performs better, but there still exist some artifacts. The results of the DenseFuse-based, U2Fusion-based, GANMcC-based, and CSF-based methods tend to be slightly unnatural (see the left close-ups in the last row). Color distortion occurred in the FusionGAN-based method, resulting in unclear fusion results. The RFN-Nest-based and PPT Fusion-based methods have limited ability to fully extract the thermal radiation information from the infrared image, leading to indistinct objects. Overall, the proposed YDTR can simultaneously preserve the complementary features from the source images, and provide pleasing fusion images with clear scenes and salient objects.


Fig. 10. Four sets of source images and their corresponding fusion results of different methods on MSRS data set. From top to bottom: infrared image, visible image, the fusion results of RP [8], CVT [9], DenseFuse [18], FusionGAN [19], U2Fusion [20], GANMcC [24], RFN-Nest [23], CSF [29], PPT Fusion [32], and the proposed YDTR. In each image, two local areas are enlarged as close-ups for better comparison.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

TABLE XI
QUANTITATIVE EVALUATION RESULTS OF NINE REPRESENTATIVE AND STATE-OF-THE-ART INFRARED AND VISIBLE IMAGE FUSION METHODS AND THE PROPOSED YDTR ON MSRS DATASET

| | $Q_{MI}$ [46] ↑ | $Q_{NCIE}$ [47] ↑ | $Q_P$ [48] ↑ | $MS\text{-}SSIM$ [49] ↑ | $Q_{CV}$ [50] ↓ |
|---|---|---|---|---|---|
| RP [8] | 0.2784 | 0.8036 | 0.2711 | 0.6989 | 124.4088 |
| CVT [9] | 0.2941 | 0.8042 | 0.3418 | 0.7293 | 54.73474 |
| DenseFuse [18] | 0.4322 | 0.8066 | **0.3981** | **0.7588** | 70.5651 |
| FusionGAN [19] | 0.3390 | 0.804 | 0.1359 | 0.7018 | 258.8455 |
| U2Fusion [20] | 0.3855 | 0.8056 | 0.3316 | 0.7456 | 70.4063 |
| GANMcC [24] | 0.4164 | 0.8059 | 0.3218 | 0.7328 | 97.5859 |
| RFN-Nest [23] | 0.3936 | 0.8058 | 0.3434 | 0.7189 | 80.9327 |
| CSF [29] | 0.3973 | 0.8054 | 0.3391 | 0.7553 | 82.7246 |
| PPT Fusion [32] | 0.3503 | 0.8043 | 0.3091 | 0.758 | 149.4944 |
| YDTR | **0.4566** | **0.8076** | 0.3800 | 0.7293 | **51.8019** |

The mean values of 361 image pairs from the MSRS dataset are listed. For each metric, the optional and suboptimal values are labeled in bold and underlined, respectively.

Table XI reports the quantitative evaluation of the nine state-of-the-art methods and the proposed YDTR. For each method, 361 testing samples are calculated, and the average score is listed. For each metric, the optimal and suboptimal values are labeled in bold and underlined, respectively. Clearly, the proposed method achieves the highest scores on $Q_{MI}$ and $Q_{NCIE}$, which demonstrates that our YDTR owns the best information preservation ability. Although YDTR ranges the suboptimal performance on $Q_P$, the margin from the best one is tiny. Therefore, the proposed method has good feature extraction capability. For $MS\text{-}SSIM$, the proposed YDTR has a modest objective performance, which illustrates that our proposed dynamic Transformer can exploit complementary features. The minimum value on $Q_{CV}$ reveals that our method is more consistent with the human visual system.

Based on the above qualitative and quantitative observations, we can draw the conclusion that our proposed YDTR outperforms other state-of-the-art methods in both visual quality and objective assessment, and exhibits pleasing generalization ability.

### F. Extension to Multi-Focus Image Fusion

To further investigate the generalization capability of our proposed YDTR, we directly apply the trained model to deal with a very different image fusion task, i.e., multi-focus image fusion, without fine-tuning. Unlike the infrared and visible images, where the different information provided by each modality is caused by various imaging techniques, the varying characteristics of multi-focus images are due to the restricted depth-of-field (DOF) of imaging devices. Specifically, the objects within the DOF remain clear while the targets outside the DOF are blurred. As the all-in-focus clear images are important to human visual perception and subsequent computer vision tasks, multi-focus image fusion can be a useful tool to produce the demanded images. In our generalization experiments, 20 image pairs from [62] are downloaded as testing samples. Four state-of-the-art multi-focus image fusion methods are adopted for comparison, which are the SFMD-based method [63], the PMGI-based method [64], the FusionDN-based method [65], and the U2Fusion-based method [20].

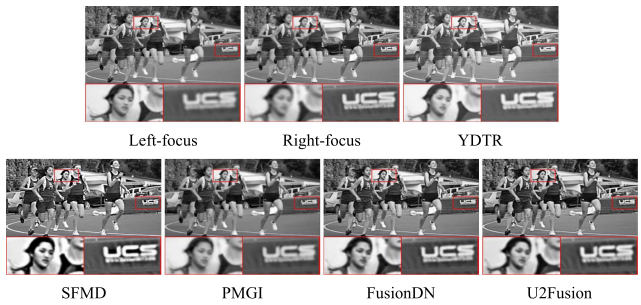


Fig. 11. A pair of multi-focus images and their corresponding fusion results of different methods. In each image, two local areas are zoomed as close-ups for better comparison.

TABLE XII
QUANTITATIVE EVALUATION RESULTS OF FOUR STATE-OF-THE-ART MULTI-FOCUS IMAGE FUSION METHODS AND THE PROPOSED YDTR

| Metrics | SFMD | PMGI | FusionDN | U2Fusion | YDTR |
|---|---|---|---|---|---|
| $Q_{MI}$ [46] ↑ | 0.6732 | 0.8614 | 0.7259 | 0.7166 | **0.9119** |
| $Q_{NCIE}$ [47] ↑ | 0.8204 | 0.8272 | 0.8221 | 0.8215 | **0.8287** |
| $Q_P$ [48] ↑ | 0.7694 | 0.771 | 0.7722 | **0.8121** | 0.7947 |
| $Q_{CV}$ [50] ↓ | 25.3231 | 5.5139 | 10.1434 | 7.7625 | **4.2887** |
| $MS\text{-}SSIM$ [49] ↑ | 0.9319 | 0.96 | 0.9513 | 0.963 | **0.9727** |

The mean values of 20 testing samples are listed. For each metric, the optional and suboptimal values are labeled in bold and underlined, respectively.

A pair of multi-focus images and their corresponding fusion results generated by different algorithms are shown in Fig. 11. It can be observed that our YDTR can produce an all-in-focus clear image. Comparatively, the SFMD-based method tends to generate unreal fusion result (see the left close-up in Fig. 11). The FusionDN-based and U2Fusion-based methods can slightly alleviate this defect but still suffer from it to a certain degree. The PMGI-based method cannot well obtain a clear enough fused image. Table XII reports the quantitative comparison on the multi-focus dataset. For each metric, the best and second-best methods are labeled in bold and underlined, respectively. Obviously, our method possesses the best objective performance on $Q_{MI}$, $Q_{NCIE}$, $Q_{CV}$, and $MS\text{-}SSIM$. Although the proposed YDTR ranges the suboptimal objective assessment on $Q_P$, the margin to the optimal is tiny. Based on the subjective and objective analyses, we can conclude that our method indeed owns promising generalization capability.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

## V. Conclusion

In this paper, we propose a novel end-to-end infrared and visible image fusion method via a Y-shape dynamic Transformer called YDTR. The proposed YDTR is composed of two Y branches for complementary information extraction and a main path for feature integration. Specifically, the two Y branches are employed to separately explore the thermal radiation information from the infrared image and texture details from the visible image. The captured features are adequately merged by the main path to further preserve important information. In addition, a dynamic Transformer module (DTRM) is designed to ensure that the proposed method can not only fully excavate local useful information but also global features. Furthermore, a loss function that consists of an SSIM term and an SF term is devised to train the proposed model in an unsupervised fashion. Ablation experiments demonstrate the effectiveness of the proposed network architecture and loss function. Subjective and objective comparisons with the other two representative traditional approaches and seven state-of-the-art algorithms illustrate the superiority and significance of the proposed YDTR. We further apply our model to address infrared and RGB-visible image fusion and multi-focus image fusion tasks without fine-tuning, and the satisfactory fusion performance demonstrates that the proposed YDTR has good generalization ability.

## References


[1] Z. Li, H. Hu, W. Zhang, S. Pu, and B. Li, "Spectrum characteristics preserved visible andnear-infrared image fusion algorithm," *IEEE Trans. Multimedia*, vol. 23, pp. 306–319, 2021.

[2] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.

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

[4] A. C. Muller and S. Narayanan, "Cognitively-engineered multisensor image fusion for military applications," *Inf. Fusion*, vol. 10, no. 2, pp. 137–149, 2009.

[5] Y. Li, H. Zhao, Z. Hu, Q. Wang, and Y. Chen, "IVFuseNet: Fusion of infrared and visible light images for depth prediction," *Inf. Fusion*, vol. 58, pp. 1–12, 2020.

[6] S. G. Kong et al., "Multiscale fusion of visible and thermal IR images for illumination-invariant face recognition," *Int. J. Comput. Vis.*, vol. 71, pp. 215–233, 2007.

[7] R. Raghavendra, B. Dorizzi, A. Rao, and G. Kumar, "Particle swarm optimization based fusion of near infrared and visible images for improved face verification," *Pattern Recognit.*, vol. 44, no. 2, pp. 401–411, 2011.

[8] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 245–253, 1989.

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

[10] M. Kumar and S. Dass, "A total variation-based algorithm for pixel-level image fusion," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2137–2143, Sep. 2009.

[11] T. Wan, N. Canagarajah, and A. Achim, "Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 624–633, Jun. 2009.

[12] H. Yin, S. Li, and L. Fang, "Simultaneous image fusion and super-resolution using sparse representation," *Inf. Fusion*, vol. 14, no. 3, pp. 229–240, 2013.

[13] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, no. 1, pp. 147–164, 2015.

[14] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 866–879, Apr. 2018.

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

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

[17] B. Wang et al., "Latent representation learning model for multi-band images fusion via low-rankand sparse embedding," *IEEE Trans. Multimedia*, vol. 23, pp. 3137–3152, 2021.

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

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

[20] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

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

[22] F. Zhao and W. Zhao, "Learning specific and general realm featurerepresentations for image fusion," *IEEE Trans. Multimedia*, vol. 23, pp. 2745–2756, 2020.

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

[24] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5005014.

[25] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, 2018, Art. no. 1850018.

[26] Z. Zhao et al., "Efficient and interpretable infrared and visible image fusion via algorithm unrolling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1186–1196, Mar. 2021.

[27] J. Li et al., "Multigrained attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5002412.

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

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

[30] V. VS, J. M. J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," 2021, *arXiv:2107.09011*.

[31] H. Zhao and R. Nie, "DNDT: Infrared and visible image fusion via DenseNet and dual-transformer," in *Proc. Int. Conf. Inf. Technol. Biomed. Eng.*, 2021, pp. 71–75.

[32] Y. Fu, T. Xu, X. Wu, and J. Kittler, "PPT fusion: Pyramid patch transformer for a case study in image fusion," 2021, *arXiv:2107.13967*.

[33] D. Rao, X. Wu, and T. Xu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," 2022, *arXiv:2201.10147*.

[34] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, 2017.

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

[36] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.

[37] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Nneural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[38] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[39] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 1–11.

[40] Z. Dai et al., "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.


Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and X. Zhai, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[42] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[43] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
[44] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[45] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8741–8750.
[46] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on information measure for performance of image fusion," *Electron. Lett.*, vol. 44, no. 18, pp. 1066–1067, 2008.
[47] Q. Wang, Y. Shen, and J. Zhang, "A nonlinear correlation measure for multivariable data set," *Physica D: Nonlinear Phenomena*, vol. 200, no. 3–4, pp. 287–295, 2005.
[48] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *Int. J. Innov. Comput. Inf. Control*, vol. 3, no. 6, pp. 1433–1447, 2007.
[49] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, 2003, vol. 2, pp. 1398–1402.
[50] H. Chen and P. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Inf. Fusion*, vol. 8, pp. 193–207, 2007.
[51] T. Xiao et al., "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 1–16.
[52] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[53] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[54] W. Tang et al., "Green fluorescent protein and phase-contrast image fusion via generative adversarial networks," *Comput. Math. Methods Med.*, vol. 2019, 2019, Art. no. 5450373.
[55] W. Tang, Y. Liu, J. Cheng, C. Li, and X. Chen, "Green fluorescent protein and phase contrast image fusion via detail preserving cross network," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 584–597, 2021.
[56] K. Lu and L. Zhang, "TBEFN: A two-branch exposure-fusion network-for low-light image enhancement," *IEEE Trans. Multimedia*, vol. 23, pp. 4093–4105, 2021.
[57] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[59] Z. Liu et al., "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.
[60] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83, pp. 79–92, 2022.
[61] W. Tang, L. Wang, and Y. Liu, "Green fluorescent protein and phase contrast image fusion via dual attention residual network," in *Proc. IEEE Int. Conf. Med. Imag. Phys. Eng.*, 2021, pp. 1–6.
[62] H. Li and X. Wu, "Multi-focus image fusion using dictionary learning and low-rank representation," in *Proc. Int. Conf. Image Graph.*, 2017, pp. 675–686.
[63] H. Li and L. Li, "Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering," *Opt. Commun.*, vol. 342, pp. 1–11, 2015.
[64] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 12797–12804.
[65] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 12484–12491.

**Wei Tang** received the B.E. degree from Wannan Medical College, Wuhu, China, in 2018, and the M.S. degree in biomedical engineering from the Hefei University of Technology, Hefei, China, in 2021. She is currently working toward the Ph.D. degree with the school of computer science, Wuhan University, Wuhan, China. Her research interests include image processing, computer vision, and information fusion.

**Fazhi He** (Member, IEEE) received the bachelor's, master's, and Ph.D. degrees from the Wuhan University of Technology, Wuhan, China. He was Postdoctoral Researcher with The State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, a Visiting Researcher with Korea Advanced Institute of Science & Technology, Daejeon, South Korea, and a Visiting Faculty Member with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He is currently a Professor with the School of Computer Science, Wuhan University, Wuhan, China. His research interests include artificial intelligence, intelligent computing, computer graphics, and image processing.

**Yu Liu** (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively. He is currently an Associate Professor with the Department of Biomedical Engineering, Hefei University of Technology, Hefei, China. His research interests include image processing, computer vision, information fusion, and machine learning. His research interests include image fusion, image restoration, visual recognition, and deep learning. Dr. Liu is the Editorial Board Member of *Information Fusion*.

Authorized licensed use limited to: Australian National University. Downloaded on May 18,2025 at 04:19:31 UTC from IEEE Xplore. Restrictions apply.